Rights, Equality and Citizenship (REC)

Programme of the European Commission

(2014-2020)

MANDOLA

Monitoring and Detecting Online Hate Speech

D1.4: Advisory Board Meeting 1 [†]

Abstract: This deliverable reports the work done in creating the MANDOLA Advisory Board and preparing its first meeting, as well as the proceedings of this meeting.

Contractual Date of Delivery	September 2016
Actual Date of Delivery	October 2016
Deliverable Security Class	Public
Editor	Nikos Frydas
Contributors	Evangelos Markatos Meltini Christodoulaki
Quality Assurance	Marios Dikaiakos

[†] This project is funded by the Rights, Equality and Citizenship (REC) Programme of the European Commission.

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
1.4.2			
1.4.1	22/10/2016	Editor	Amendments / corrections from AB members
1.4.0	11/10/2016	Editor	First draft.

Table of Contents

1 Introduction

This document, comprises the following chapters:

1. **Chapter 2** [*Aims & Objectives of the Advisory Board (AB)*]: This chapter describes the aims & objectives of the Advisory Board, as well as the practical constraints taken into consideration, when examining candidates.
2. **Chapter 3** [*Methodology used to populate the AB*]: This chapter describes the methodology used to populate the Advisory Board.
3. **Chapter 4** [*Proceedings of the AB1*]: This chapter gives the proceedings of the Advisory Board Meeting 1 (**AB1**).
4. **Chapter 5** [*Conclusions & Lessons Learned*]: This chapter gives the conclusions and lessons learned from AB1.
5. The document includes the following **appendices**:
   a. Appendix A: Agenda of AB1 (Advisory Board Meeting 1)
   b. Appendix B: Advisory Board presentation
   c. Appendix C: Introduction to MANDOLA presentation
   d. Appendix D: Technical Infrastructure presentation
   e. Appendix E: Definition of Hate Speech & Legal Framework presentation
   f. Appendix F: Brainstorming Panel I / Question 1
   g. Appendix G: Brainstorming Panel I / Question 2
   h. Appendix H: Brainstorming Panel I / Question 3
   i. Appendix I: Brainstorming Panel I / Question 4
   j. Appendix J: Brainstorming Panel II / Question 1
   k. Appendix K: Brainstorming Panel II / Question 2
   l. Appendix L: Brainstorming Panel II / Question 3
   m. Appendix M: Brainstorming Panel II / Question 4

2 Aims & Objectives of the Advisory Board (AB)

This chapter describes the aims & objectives of the Advisory Board, as well as the practical constraints taken into consideration.

The ***aim*** of the task undertaken it to compose the ***optimum AB***, under the ***practical constraints*** of the project.

The Chapter comprises the following sections:

1. The Objectives of the MANDOLA AB
2. AB Constraints
3. AB Membership

2.1 The Objectives of the MANDOLA AB

Setting up an Advisory Board “that will steer this project” is the goal of WS1.3. The delivery of the following outputs is part of the project’s contractual obligations:

1. D1.4 Advisory Board Meeting 1 Target group: ALL
2. D1.5 Advisory Board Meeting 2 Target group: ALL

The current document constitutes deliverable D1.4.

2.1.1 AB duties in general

In general, an Advisory Board provides non-binding strategic advice. Among the reasons for creating an AB are the following:

- Seek expertise outside MANDOLA.
- Complement existing strengths.
- Counsel on issues raised by MANDOLA.
- Become a resource for MANDOLA managers.
- Provide un-biased ideas.
- Monitor project performance.

2.1.2 AB duties in particular

According to the MANDOLA project objectives, the Advisory Board should have the following characteristics:

- AB will **steer** the project.
- AB will help **spread** the project message well **beyond** participant Member States.
- AB will assist the **promotion** of the developed technologies and tools.
- AB will provide valuable **feedback** & **market guidelines** on progress & results.
- AB will further **enhance** impact & **dissemination** of MANDOLA’s ideas.
- AB will foster dialogue & **debate**.
- AB will serve as a source of **expertise**.

2.2 AB Constraints

Project constraints place an upper limit of **20** to the number of external AB members who reside outside Brussels. In addition, the **AB members must be EU residents**.

The meeting room made available has a capacity of 25. This implies that with a total of nine internal AB members, the **external AB members should be restricted to 16**.

> MANDOLA project partners are grateful to the ***European Office of Cyprus***, in Rue du Luxembourg 3, Brussels, who made their meeting room available, free of charge.

2.3 AB Membership

In general, AB members must be individuals

1. with personal qualities and
2. representing an *important* entity, where *important* is understood to mean *important for the project*, and
3. with knowledge of the issues the project deals with and
4. with good command of English and
5. with the ability to be present at the AB meetings in Brussels.

Given the above and the project objectives (see «The Objectives of the MANDOLA AB», above), AB members shall then be drawn from:

- Academia
- NGOs
- LEA
- Internet Industry
- Government
- other

3 Methodology used to populate the AB

This chapter describes the methodology used to populate the Advisory Board.

It was decided to follow the methodology below:

1. Create a super-list of 50-60 individuals, candidates for the AB.
2. Assess the suitability of each individual across a number of attributes.
3. Combine the marks/attribute into an overall score/individual.
4. Order the individuals according to their score.
5. Invite the top 16 individuals.
6. Once an individual accepts an invitation, the individual is moved to the top of the list.
7. Once an individual declines the invitation, the individual is moved to the bottom of the list.
8. Continue until you have 16 acceptances.

3.1 Attributes of AB candidates

An optimum AB, would be one which would satisfy the aims and objectives discussed under Chapter 2 (see p. 7). The basic qualities required, from the AB, are:

1. A focused range of expertise:
    a. Child Care
    b. Cybersecurity (Leg)
    c. Cybersecurity (Tech)
    d. Hate Speech
    e. Hotline
    f. Human Rights
    g. Linguistics
2. Balanced representation of AB members' organizations:
    a. Academia
    b. Industry
    c. Intl ORG
    d. LEA
    e. Mass Media
    f. NGO
    g. State
3. Wide and balanced representation of nationalities.
4. A balanced gender composition.

In addition to the above, it was thought appropriate to balance MANDOLA members[1] (the **MEMBERS**) recommendations, as well as MANDOLA member organizations' recommendations.

Finally, given that the total AB cost depends mainly on travelling expenses, it was thought that the distribution of the areas of residence should also be balanced.

Given the above, the **ATTRIBUTES** for assessing AB candidates are:

[1] The team of individuals who participate in the MANDOLA project, on behalf of the Coordinator and the Principal Contractors.

A. Primary Area of Expertise [of the AB member]
B. Type of Organization [to which the AB member belongs]
C. [MANDOLA] Member [proposing an AB candidate]
D. [MANDOLA] Member Organization [proposing an AB candidate]
E. Nationality [of the AB candidate]
F. Area of Residence [of the AB candidate]
G. Gender [of the AB candidate]

3.2 Population of the Super-List

Super-List is the list of all individuals, considered for AB participation.

The Super-List was populated via recommendations from MEMBERS, who were invited on 5/11/2015 to fill a suitable recommendation form. This process lasted for almost seven months and was interactive.

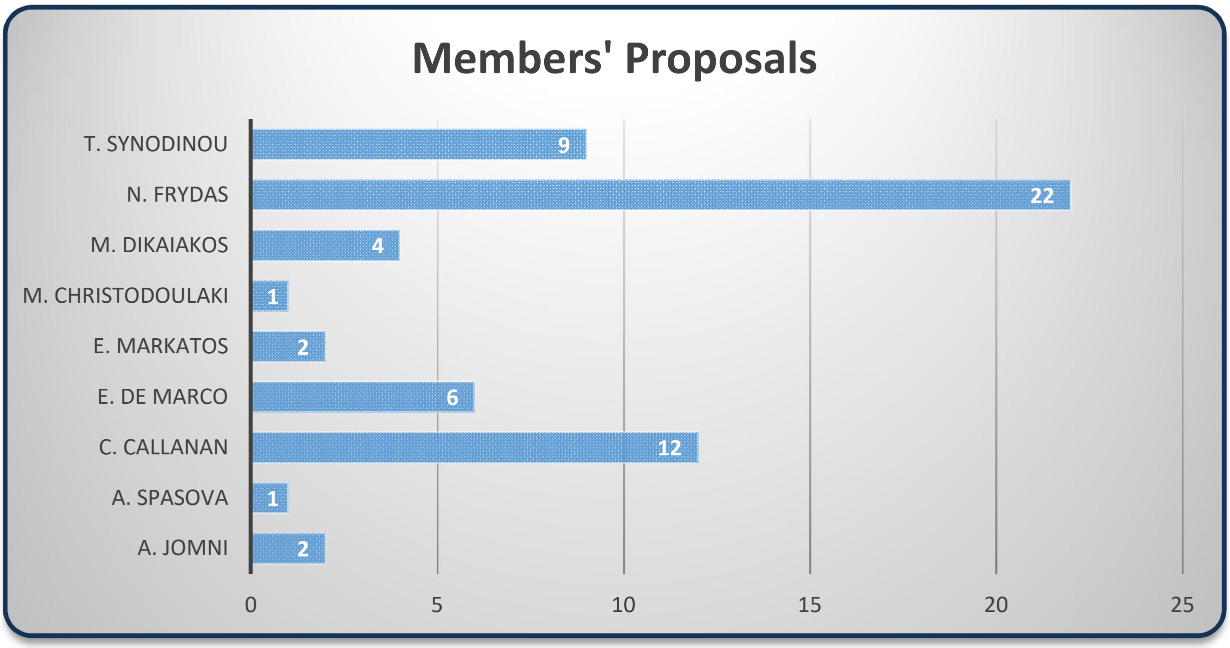


Figure 1: MANDOLA Members' recommendations regarding AB composition.

The results are depicted in Figure 1 and in Figure 2.

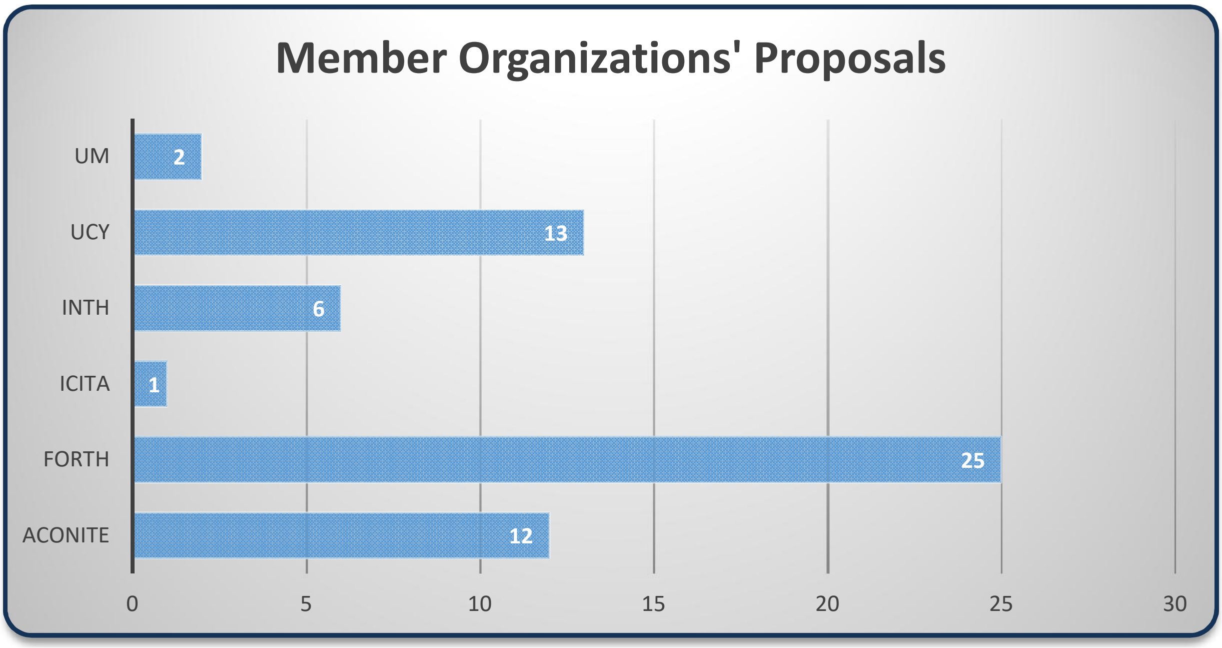


Figure 2: MANDOLA Member Organizations' recommendations regarding AB composition.

The Super-List composition regarding the Type of Organization the AB candidates belong to, is depicted in Figure 3.

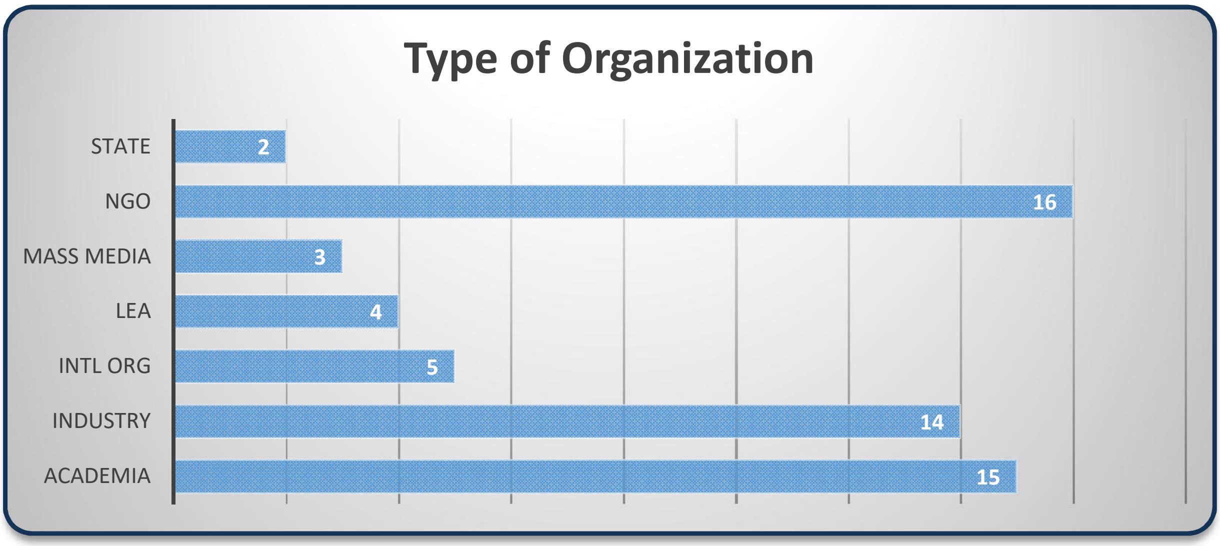


Figure 3: Type of Organization AB candidates belong

The Super-List composition regarding the Primary Area of Expertise of the AB candidates, is depicted in Figure 4.

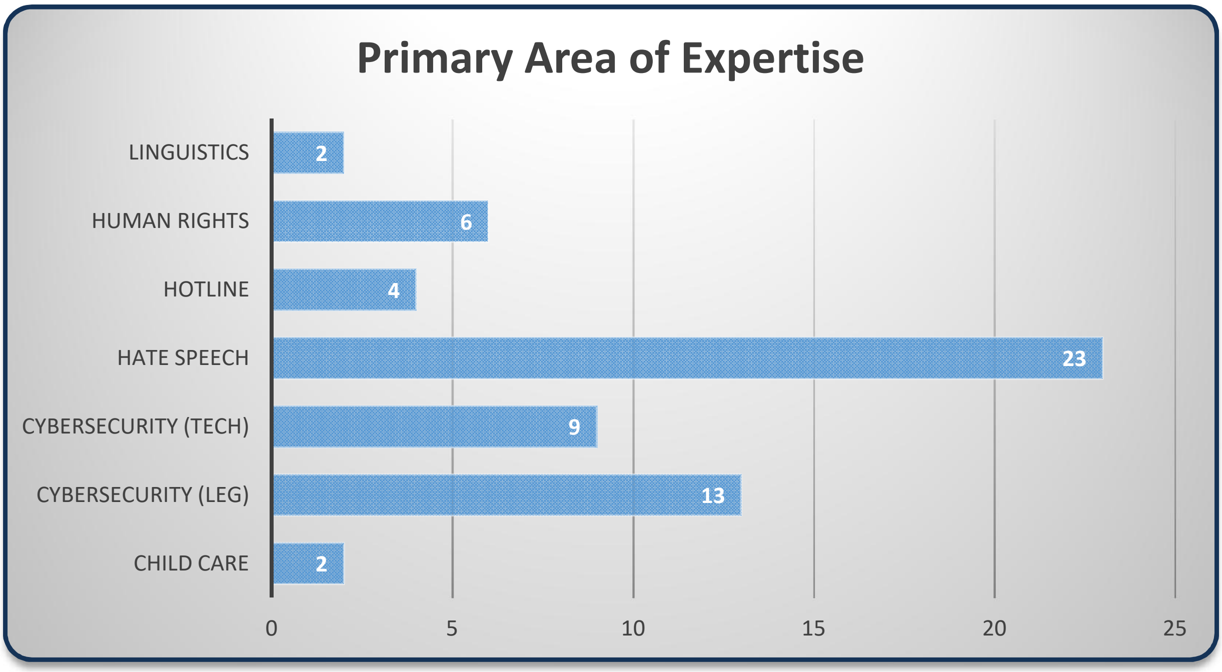


Figure 4: Primary Area of Expertise of AB candidates

The Super-List composition regarding the gender distribution of the AB candidates, is depicted in Figure 5. "X" denotes unknown gender, as it was not known which individual would represent the candidate organization.

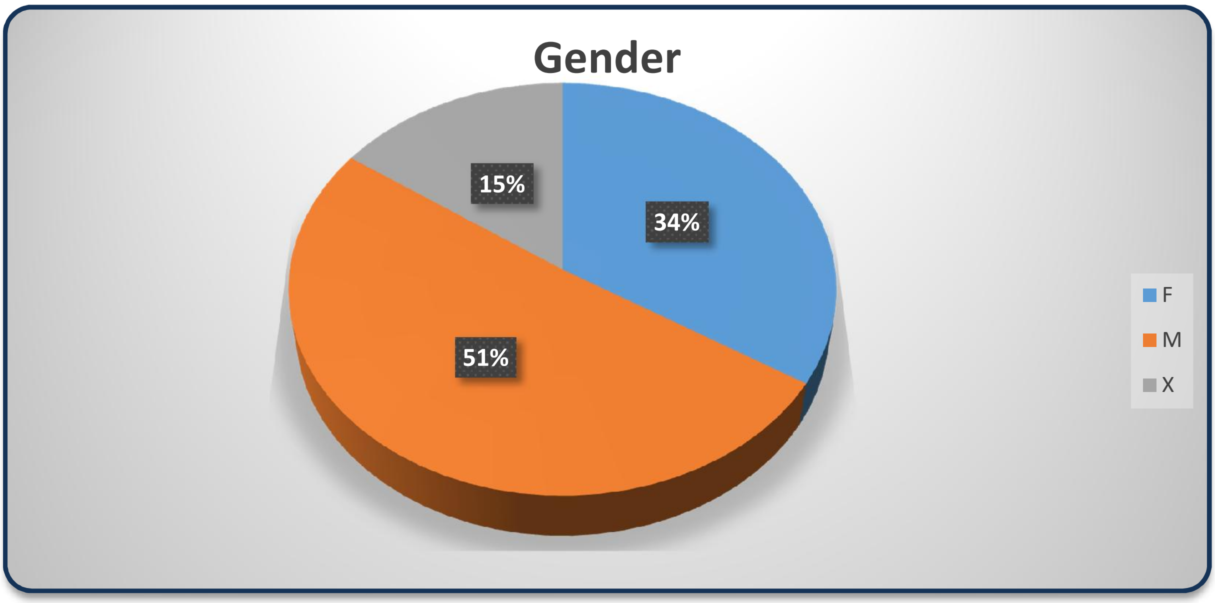


Figure 5: Gender distribution of AB candidates.

The Super-List composition regarding the Nationality of the AB candidates, is depicted in Figure 6. "European" denotes unknown nationality, as it was not known which individual would represent the candidate organization.

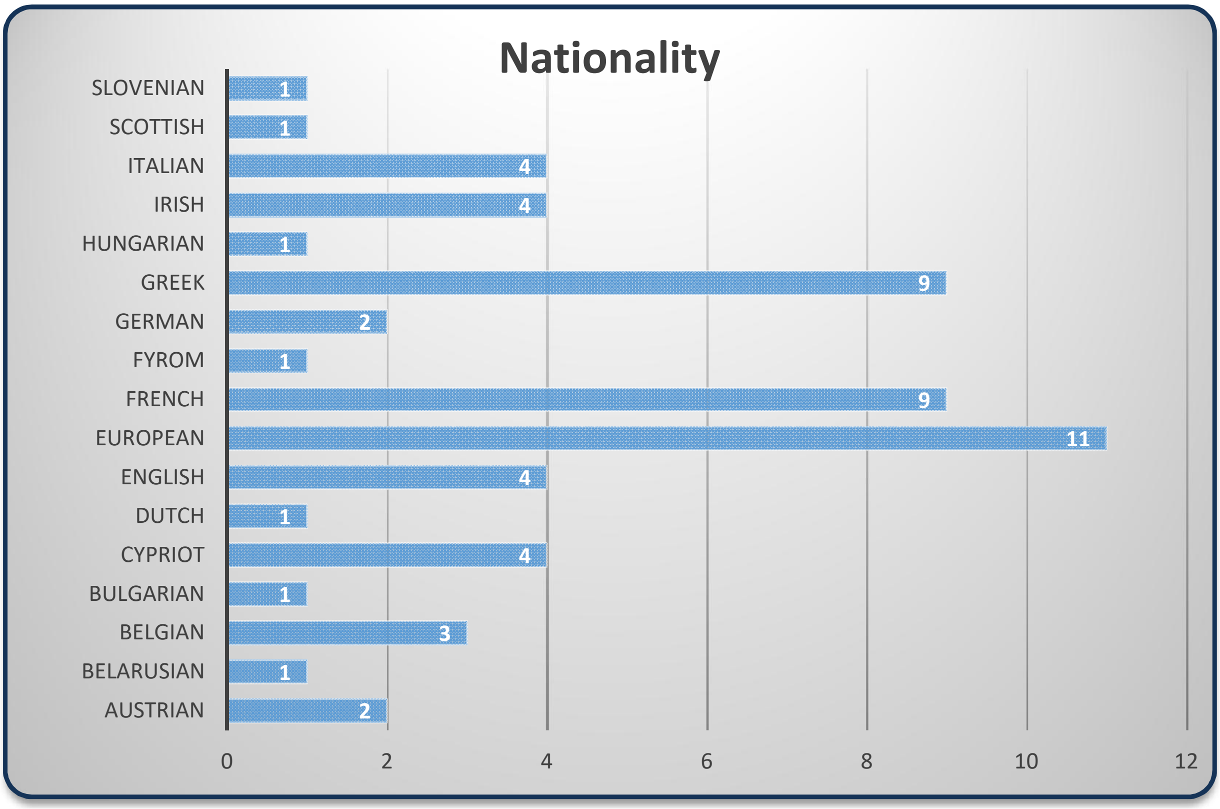


Figure 6: Nationality of AB candidates.

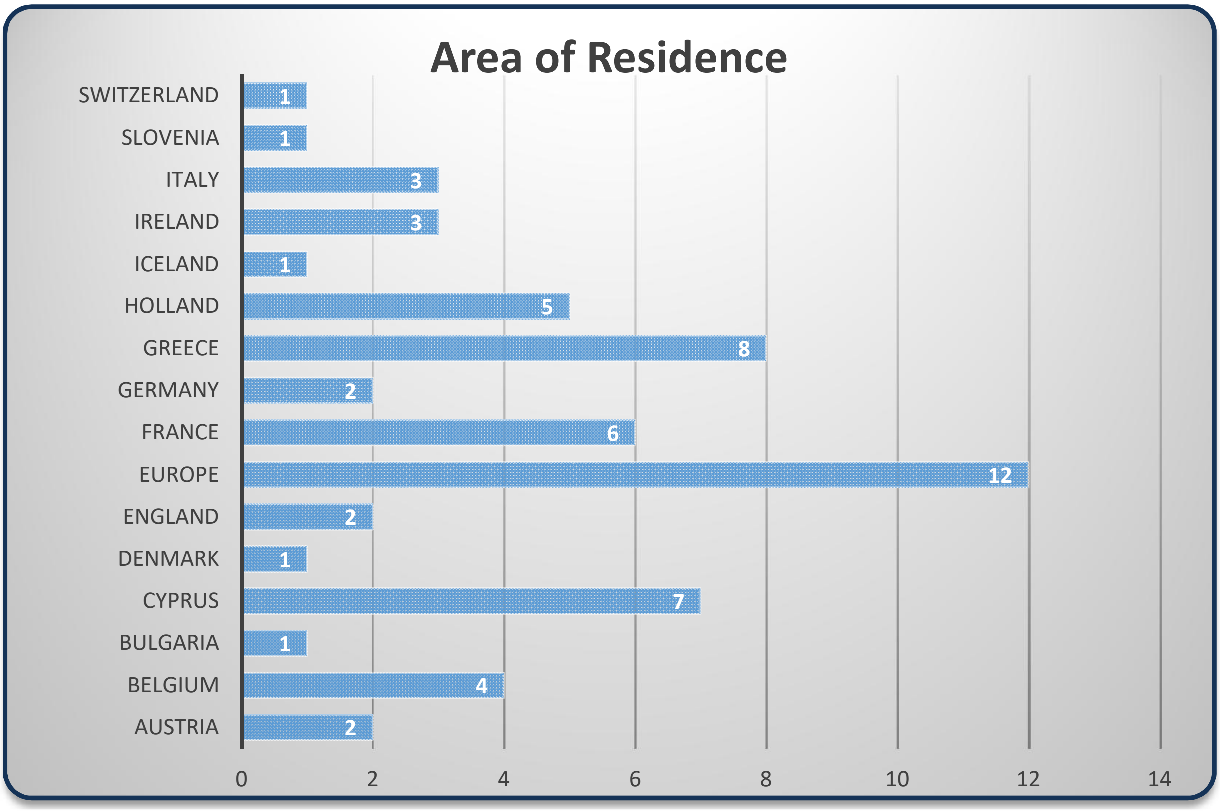


Figure 7: Area of Residence of AB candidates.

3.3 Candidate Score

Each candidate is awarded a mark for each of the ATTRIBUTES introduced earlier on:

A. Primary Area of Expertise [of the AB member]
B. Type of Organization [to which the AB member belongs]
C. [MANDOLA] Member [proposing an AB candidate]
D. [MANDOLA] Member Organization [proposing an AB candidate]
E. Nationality [of the AB candidate]
F. Area of Residence [of the AB candidate]
G. Gender [of the AB candidate]

The individual marks, one per ATTRIBUTE per candidate are weighted and summed to produce the candidate score. The candidate score is thus the weighted average of the candidates' marks per ATTRIBUTE.

3.3.1 Primary Area of Expertise [of the AB member]

As discussed in §3.1, above, the following expertise is considered desirable for the AB:

Table 1: Primary Area of Expertise of AB Members and Weighting

#	Primary Area of Expertise	W-Ex	W-Ex%
01	Child Care	**5**	*11%*
02	Cybersecurity (Leg)	**7**	*15%*
03	Cybersecurity (Tech)	**7**	*15%*
04	Hate Speech	**10**	*21%*
05	Hotline	**9**	*19%*
06	Human Rights	**8**	*17%*
07	Linguistics	**1**	*2%*

Each entry is given a weight from 1 to 10, according to importance (field "W-Ex"). Field "W-Ex%", indicates the % value of the weights. For example, "Hate Speech" is the most important Expertise and is, thus, given a weight of 10, etc.

Each candidate is given a mark from 1 to 10, as following: All candidates with the same Expertise, say "Human Rights", are sorted in order of suitability for the AB, with the most suitable scoring 10, the next most suitable 9, etc.

3.3.2 Type of Organization [to which the AB member belongs]

As discussed in §3.1, above, the following Types of Organization are considered desirable for the AB:

Table 2: Types of Organization of AB Members and Weighting

#	Type of Organization	W-Or	W-Or%
01	Academia	**5**	*9%*
02	Industry	**10**	*18%*
03	Intl ORG	**8**	*14%*
04	LEA	**9**	*16%*
05	Mass Media	**9**	*16%*
06	NGO	**7**	*12%*
07	State	**9**	*16%*

Each entry is given a weight from 1 to 10, according to importance (field “W-Or”). Field “W-Or%”, indicates the % value of the weights. For example, “Industry” is considered the most important Type of Organization and is, thus, given a weight of 10, etc.

Each candidate is given a mark from 1 to 10, as following: All candidates with the same Type of Organization, say “NGO”, are sorted in order of suitability for the AB, with the most suitable scoring 10, the next most suitable 9, etc.

3.3.3 [MANDOLA] Member [proposing an AB candidate]

Entries are the MEMBERS. Each entry is given the same weight.

Each candidate is given a mark from 1 to 10, as following: All candidates proposed by the same MEMBERS, say N. Frydas, are sorted in order of suitability for the AB, with the most suitable scoring 10, the next most suitable 9, etc.

3.3.4 [MANDOLA] Member Organization [proposing an AB candidate]

Entries are the MEMBERS’ Organizations. Each entry is given the same weight.

Each candidate is given a mark from 1 to 10, as following: All candidates proposed by the same MEMBERS’ Organization, say FORTH, are sorted in order of suitability for the AB, with the most suitable scoring 10, the next most suitable 9, etc.

3.3.5 Nationality [of the AB candidate]

Entries are the AB candidate’s Nationality. Each entry is given the same weight.

Each candidate is given a mark from 1 to 10, as following: All candidates of the same Nationality, say Greek, are sorted in order of suitability for the AB, with the most suitable scoring 10, the next most suitable 9, etc.

3.3.6 Area of Residence [of the AB candidate]

Entries are the Areas of Residence of the AB candidates (e.g., Austria, Belgium, Bulgaria, Cyprus, etc.). Each entry is given different weights, according to distance from Brussels. Brussels get the top mark, 10, while Cyprus get 5, etc.

Each candidate is given a mark from 1 to 10, as following: All candidates from the same Area of Residence, say Greece, are sorted in order of suitability for the AB, with the most suitable scoring 10, the next most suitable 9, etc.

3.3.7 Gender [of the AB candidate]

Entries are “Male”, “Female” & “X” (unknown yet). “Female” is given the top mark, 10, as there are less female candidates in the Super-List. “Male” is given 5 and “X” 7 (the average between “F” & “M”).

Each candidate is given a mark from 1 to 10, as following: All, say, male candidates are sorted in order of suitability for the AB, with the most suitable scoring 10, the next most suitable 9, etc.

3.4 Prioritizing AB Candidates

Each AB candidate obtains a certain score, as described previously. According to this score, the candidate is awarded a status, and if appropriate, an invitation, or an enquiry is sent. Status is modified according to the candidate’s response.

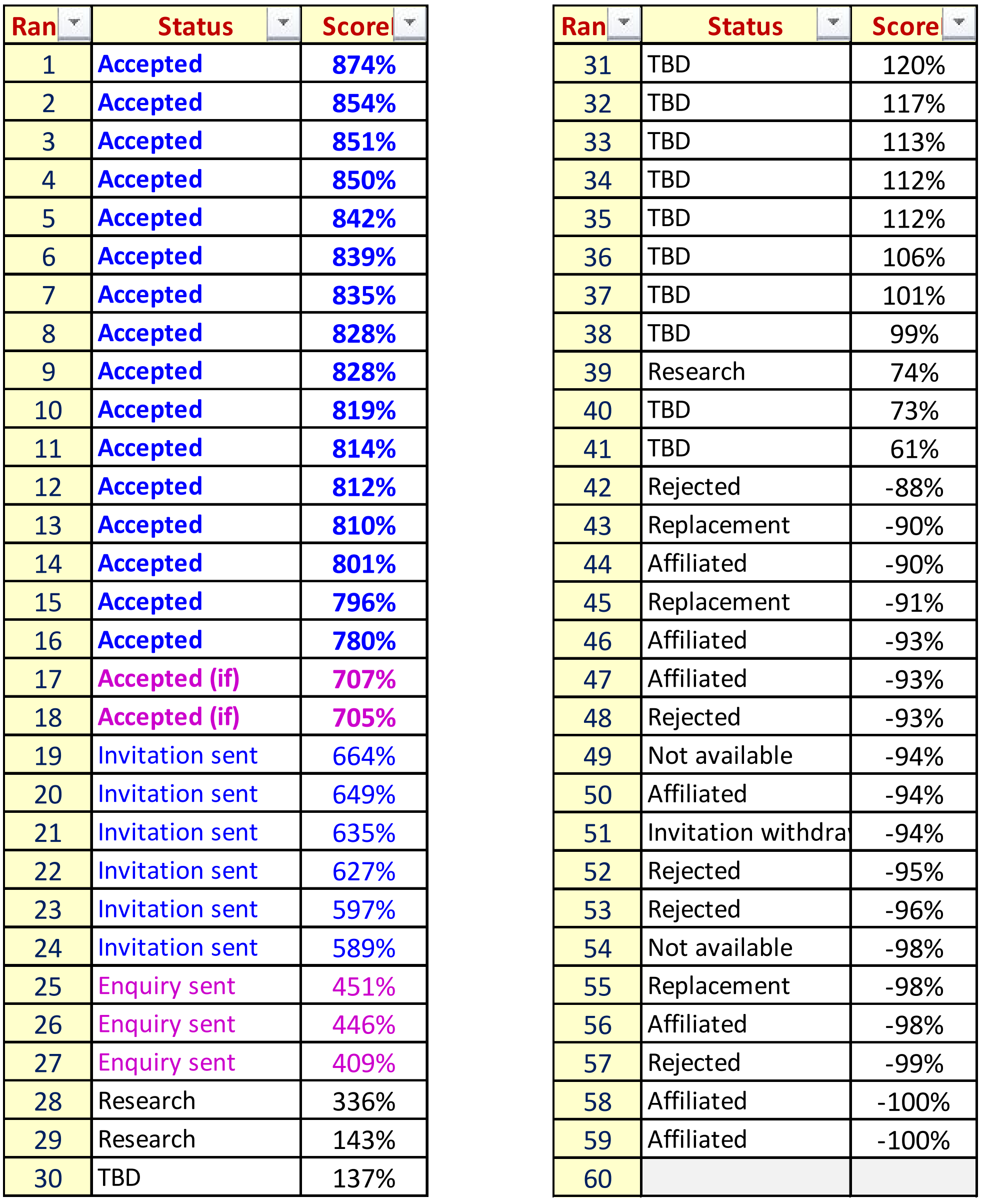
Status takes on the following values:

1. Accepted: The candidate has accepted the invitation. The system awards extra marks, so that the candidates who accepted move on to the top of the list. The entry is coloured **bold blue**.
2. Accepted (if): The candidate has accepted the invitation, under certain conditions. The system awards extra marks, less than for “accepted”, so that these candidates move on the list, just below the previous category. The entry is coloured **bold violet**.
3. Affiliated[2]: The candidate has been sent an invitation, wants to be a member of the AB but is not able to be physically present at AB1. Affiliated members will be sent the material of the AB1.
4. Enquiry sent: This is an exploratory letter, just short of an invitation. The entry is coloured violet.
5. Invitation 2b sent: This is a candidate to whom an invitation will be sent.
6. Invitation sent: This is a candidate to whom an invitation has been sent. The system awards them extra marks, more than for the category “Enquiry sent”, but less than for the category “Accepted (if)”. The entry is coloured blue.
7. Invitation withdrawn[2]: This is a candidate to whom an invitation has been sent, but then withdrawn, in writing, because there was no response after three reminders.
8. Not available[2]: This is a candidate who has changed employment.
9. Rejected[2]: This candidate has declined the invitation.
10. Replacement[2]: This candidate has recommended a colleague in his/her position.
11. Research: This candidate is researched with the aim of possibly changing status to “Invitation 2b sent”.
12. TBD: This candidate is researched with the aim of possibly changing status to “Research”.

[2] The system subtracts marks, so that such candidates move to the bottom of the list.

The status of the Super-List, at the end of this process, is shown in Table 3 (see p. 17).

Table 3: AB Candidates ranked in order of decreasing score, as of Aug. 2016.



Ran	Status	Score
1	**Accepted**	**874%**
2	**Accepted**	**854%**
3	**Accepted**	**851%**
4	**Accepted**	**850%**
5	**Accepted**	**842%**
6	**Accepted**	**839%**
7	**Accepted**	**835%**
8	**Accepted**	**828%**
9	**Accepted**	**828%**
10	**Accepted**	**819%**
11	**Accepted**	**814%**
12	**Accepted**	**812%**
13	**Accepted**	**810%**
14	**Accepted**	**801%**
15	**Accepted**	**796%**
16	**Accepted**	**780%**
17	**Accepted (if)**	**707%**
18	**Accepted (if)**	**705%**
19	Invitation sent	664%
20	Invitation sent	649%
21	Invitation sent	635%
22	Invitation sent	627%
23	Invitation sent	597%
24	Invitation sent	589%
25	Enquiry sent	451%
26	Enquiry sent	446%
27	Enquiry sent	409%
28	Research	336%
29	Research	143%
30	TBD	137%
31	TBD	120%
32	TBD	117%
33	TBD	113%
34	TBD	112%
35	TBD	112%
36	TBD	106%
37	TBD	101%
38	TBD	99%
39	Research	74%
40	TBD	73%
41	TBD	61%
42	Rejected	-88%
43	Replacement	-90%
44	Affiliated	-90%
45	Replacement	-91%
46	Affiliated	-93%
47	Affiliated	-93%
48	Rejected	-93%
49	Not available	-94%
50	Affiliated	-94%
51	Invitation withdra	-94%
52	Rejected	-95%
53	Rejected	-96%
54	Not available	-98%
55	Replacement	-98%
56	Affiliated	-98%
57	Rejected	-99%
58	Affiliated	-100%
59	Affiliated	-100%
60		

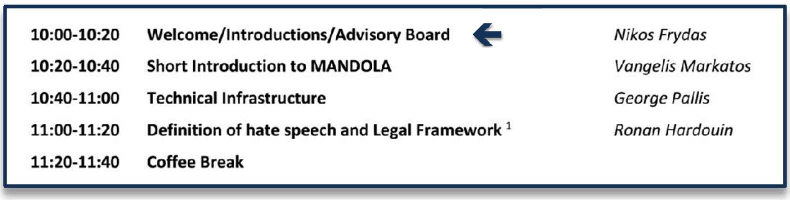
As a result of the above procedure the 16 external members of the Advisory Board were finally selected.

4 Proceedings of the AB1

This chapter gives the proceedings of the Advisory Board Meeting 1 (AB1). The chapter will be partitioned into the AB1 Agenda items (see "Appendix A: Agenda of AB1 (Advisory Board Meeting 1)"Appendix A: Agenda of AB1 (Advisory Board Meeting 1).

4.1 Welcome/Introduction/Advisory Board

Nikos Frydas welcomed the AB members and went on to present briefly the procedure by which the AB external members were selected. For the presentation see *Appendix B: Advisory Board presentation*, in p. 36.



10:00-10:20	Welcome/Introductions/Advisory Board ←	*Nikos Frydas*
10:20-10:40	Short Introduction to MANDOLA	*Vangelis Markatos*
10:40-11:00	Technical Infrastructure	*George Pallis*
11:00-11:20	Definition of hate speech and Legal Framework [1]	*Ronan Hardouin*
11:20-11:40	Coffee Break	

Following that, each AB member introduced him/her-self.

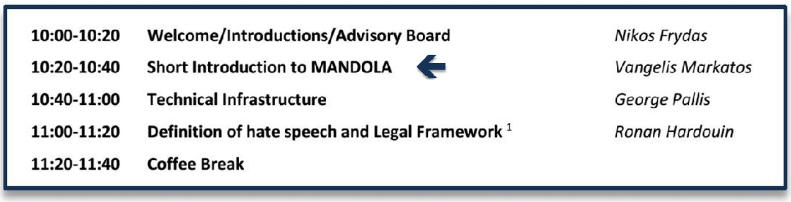
AB Internal Members:

1.	Albena	Spasova	ICITA
2.	Alvaro	Ortigosa	UAM
3.	Christian	Castane	UM1
4.	Cormac	Callanan	ACONITE
5.	Evangelos	Markatos	FORTH
6.	George	Pallis	UCY
7.	Meltini	Christodoulaki	FORTH
8.	Nikos	Frydas	FORTH
9.	Ronan	Hardouin	INTHEMIS

NOTE: The names of the 16 external AB members are currently withheld.

4.2 Short Introduction to MANDOLA

Evangelos Markatos from FORTH, the project leader, made a short introduction to the MANDOLA project activities:



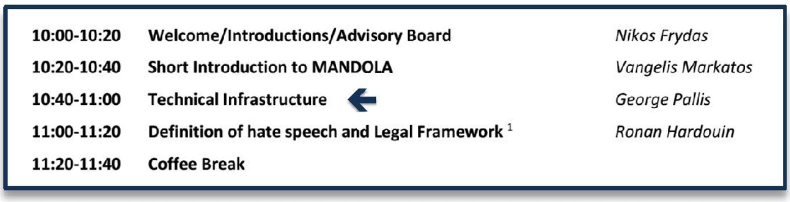
10:00-10:20	Welcome/Introductions/Advisory Board	*Nikos Frydas*
10:20-10:40	Short Introduction to MANDOLA ←	*Vangelis Markatos*
10:40-11:00	Technical Infrastructure	*George Pallis*
11:00-11:20	Definition of hate speech and Legal Framework [1]	*Ronan Hardouin*
11:20-11:40	Coffee Break	

1. Monitoring the on-line hate speech in the EU. The importance of this activity is nicely determined by Lord Kelvin: "If you cannot measure it you cannot improve it".
2. Frequently asked questions.
3. Legal:
   a. What is hate speech?
   b. Legal framework in EU member states.

For the presentation see *Appendix C: Introduction to MANDOLA presentation*, in p. 38.

4.3 Technical Infrastructure

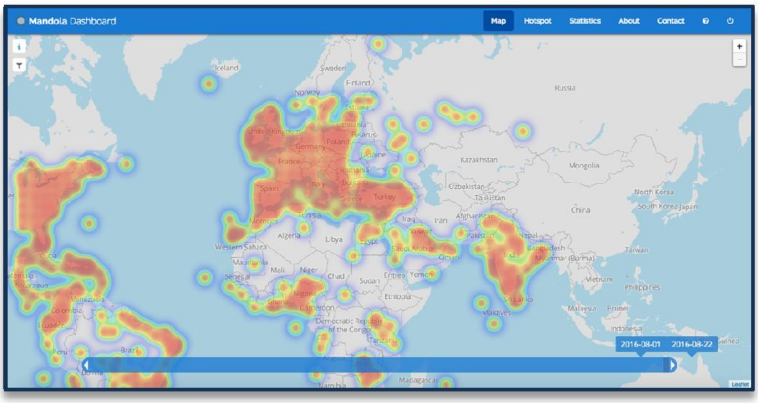
George Pallis from UCY, described the technical infrastructure needed to monitor the spread and penetration of on-line hate-related speech, as well as the necessary reporting tools that will connect citizens with the police.



10:00-10:20	Welcome/Introductions/Advisory Board	*Nikos Frydas*
10:20-10:40	Short Introduction to MANDOLA	*Vangelis Markatos*
10:40-11:00	Technical Infrastructure ←	*George Pallis*
11:00-11:20	Definition of hate speech and Legal Framework [1]	*Ronan Hardouin*
11:20-11:40	Coffee Break	

The presentation briefly referred to the following:

1. Monitoring Dashboard
2. Reporting Portal
3. Data Collection & Processing
4. Monitoring Dashboard Architecture
5. Data Analysis
6. Multi-lingual Corpus
7. Social Scientists
8. Smartphone app



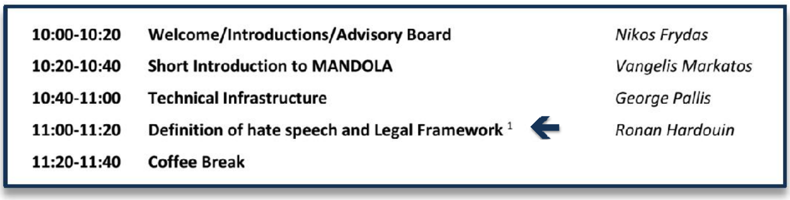
The presentation gave rise to an interesting discussion. Issues discussed include the following:

- Importance and difficulty of measuring tweets & websites.
- Importance of the distinction of the culture and the language.
- Recent progress of artificial intelligence.
- Change of the meaning of words with times.
- Project database use for other categories of illegal content.

For the presentation see *Appendix D: Technical Infrastructure presentation*, in p. 41.

4.4 Definition of Hate Speech & Legal Framework

Ronan Hardouin, from INTHEMIS, described the work done so far on the Definition of Hate Speech. In particular, a great amount of work has been done on a comparative analysis of the following EU states:



10:00-10:20	Welcome/Introductions/Advisory Board	*Nikos Frydas*
10:20-10:40	Short Introduction to MANDOLA	*Vangelis Markatos*
10:40-11:00	Technical Infrastructure	*George Pallis*
11:00-11:20	Definition of hate speech and Legal Framework [1] ←	*Ronan Hardouin*
11:20-11:40	Coffee Break	

1. Belgium
2. Bulgaria
3. Cyprus
4. France
5. Germany
6. Greece
7. Ireland

8. Netherlands
9. Romania
10. Spain

… for which 18 “potentially” illegal behaviours were identified.

Among the findings the following are included:

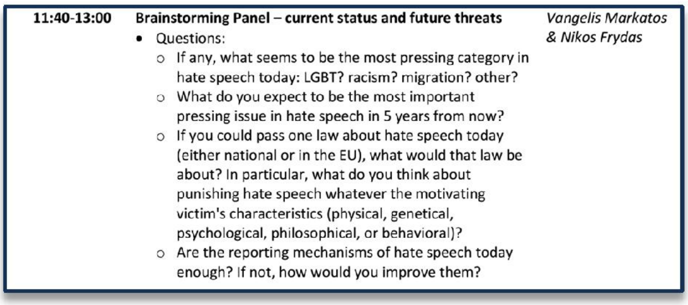
- Important disparities between legislations.
- Lack of proper transpositions of International and European legal instruments.
- Coexistence, at the domestic levels, between different provisions targeting close behaviours.

The presentation gave rise to an interesting discussion, which focused on the definition of hate speech, in general, and in particular in categorizing web content and tweets.

For the presentation see *Appendix E: Definition of Hate Speech & Legal Framework presentation*, in p. 44.

4.5 Brainstorming Panel I

In this session, four questions were given to the AB. For each question, the members wrote their answers on sticky notes, which were then collected, read, displayed on the wall and recorder for processing.

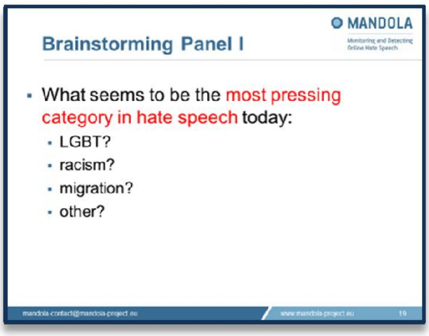


11:40-13:00 Brainstorming Panel – current status and future threats *Vangelis Markatos & Nikos Frydas*

- Questions:
  - If any, what seems to be the most pressing category in hate speech today: LGBT? racism? migration? other?
  - What do you expect to be the most important pressing issue in hate speech in 5 years from now?
  - If you could pass one law about hate speech today (either national or in the EU), what would that law be about? In particular, what do you think about punishing hate speech whatever the motivating victim's characteristics (physical, genetical, psychological, philosophical, or behavioral)?
  - Are the reporting mechanisms of hate speech today enough? If not, how would you improve them?

4.5.1 Panel I / Question 1

Question: "*What seems to be the most pressing category in hate speech today: LGBT?, racism?, migration?, or other?*".



Brainstorming Panel I

MANDOLA Monitoring and Detecting Online Hate Speech

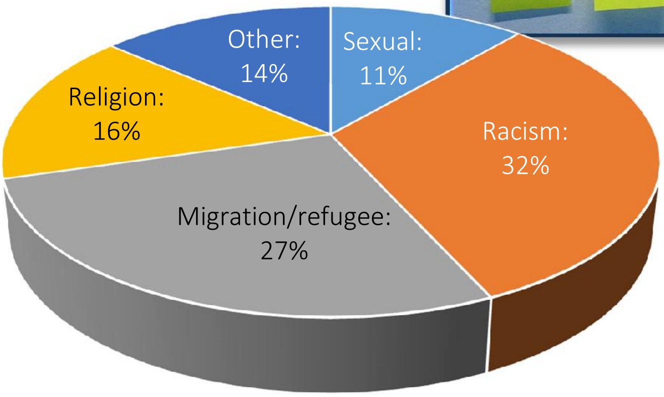
- What seems to be the most pressing category in hate speech today:
  - LGBT?
  - racism?
  - migration?
  - other?

mandola-contact@mandola-project.eu www.mandola-project.eu 19

Answers: [3]

1. LGBT: ½ + ½ = 1
2. Racism: 1 + ½ + 1 + ½ + 1 + 1 + 1 + 1 = 7
3. Migration: ½ + 1 + 1 + ½ + ½ + 1 = 4½
4. Depends on the region. CEE: LGBT [½], W. Europe: Migration [½] = 1
5. Anti-Islamic: ½
6. Refugee / asylum seeker: ½ + ½ = 1
7. Religion: ½ + 1 + ½ + ½ = 2½
8. Most pressing. Need to understand freedom & responsibility: 1
9. Other: 1
10. The general l?ck of a ??????[4] ~ what constitutes hate speech: 1
11. Sexual stereotypes: 1
12. Antisemitism: ½

The above findings may be grouped as following:



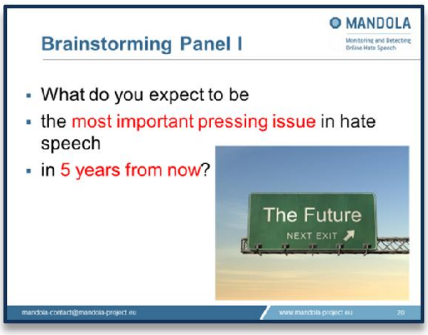
[3] Every member has one 'vote'. Hence, if a member gives n answers (n=1,2,…) to a question, then each of the member's answers carries a weight of 1/n.

[4] Not clear

A. Sexual: 2½
B. Racism: 7
C. Migration/refugee: 6
D. Religion: 3½
E. Other: 3

4.5.2 Panel I / Question 2

Question: “*What do you expect to be the most important pressing issue in hate speech in 5 years from now?*”.



MANDOLA
Monitoring and Detecting Online Hate Speech

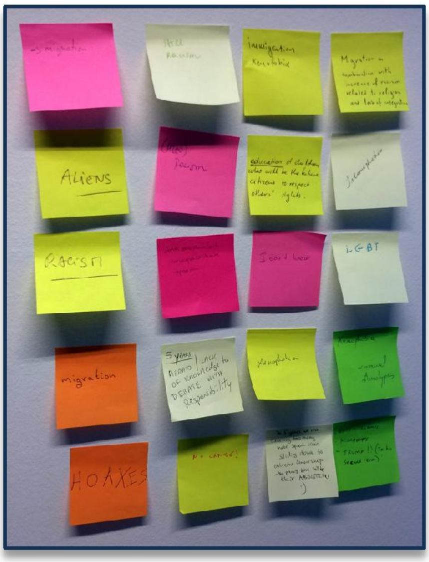
Brainstorming Panel I

- What do you expect to be
- the most important pressing issue in hate speech
- in 5 years from now?

mandola.contact@mandola-project.eu
www.mandola-project.eu
20

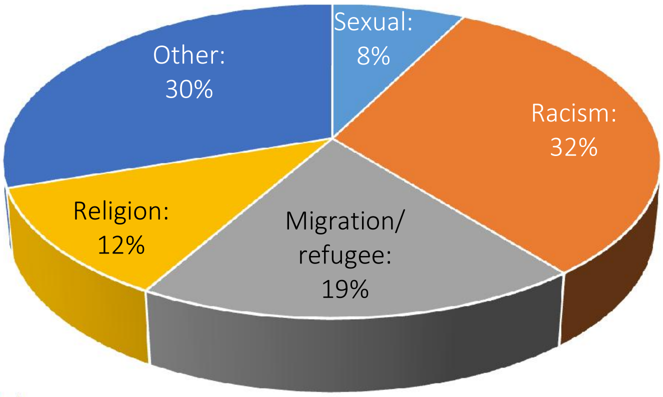
Answers:

1. Migration: 1 + 1 +⅓ + ½ + ½ = 3⅓
2. Racism: 1 + 1 + 1 = 3
3. Xenophobia: ½ + 1 + ½ = 2
4. Migration [½] in combination with increase of racism related to religion [½] and lack of integration: 1
5. Aliens: 1
6. Education of children who will be the future citizens to respect other’s rights: 1
7. Islamophobia: 1 + ½ + ⅓ = 1⅚
8. I don’t know: 1
9. LGBT: 1
10. Afraid lack of knowledge to debate with responsibility: 1
11. Sexual stereotypes: ½
12. Hoaxes: 1
13. No change: 1
14. We risk creating too many hate speech laws sliding down to extreme censorship – the pressing need is their abolition ☺: 1
15. Trump: ⅓



The above findings may be grouped as following:

A. Sexual: 1½
B. Racism/Xenophobia: 6⅓
C. Migration/refugee: 3⅚
D. Religion: 2⅔
E. Other: 6



Comparing the findings of this Question, with those of Question 1, the following findings emerge:

A. Sexual: Down from 11% to 8% (27%↘)
B. Racism/Xenophobia: No change (32%)
C. Migration/refugee: Down from 27% to 19% (30%↘)
D. Religion: Down from 16% to 12% (25%↘)
E. Other: Up from 14% to 30% (133%↗)

4.5.3 Panel I / Question 3

Question: “*If you could pass one law about hate speech today (either national or in the EU), what would that law be about?*”.

MANDOLA Monitoring and Detecting Online Hate Speech

Brainstorming Panel I

- If you could pass one law about hate speech today (either national or in the EU), what would that law be about?

mandola-contact@mandola-project.eu www.mandola-project.eu 21

Answers:

1. EU law equivalent to US 1st Amendment to clarify boundary of free speech.
2. A law about religion (freedom & tolerance).
3. We don’t need any more laws, we need other solutions.
4. Laws are not always the solution. More funding for research and education.
5. Humour law to avoid humour being considered as hate speech.
6. Start with building mechanisms to IMPLEMENT properly and fully whatever legislation we have. Enforce better cooperation with authorities upon providers (F/B, YouTube).
7. I would pass the anti-hate speech law. I’m still waiting for mandola’s definition of hate speech, though ☺.
8. Law about regulation. ISPs to have more responsibility in controlling content and for social media to have the obligation to enhance reporting mechanisms.
9. Give equal rights to LGBT.
10. Not allow social media to publish hate speech content.
11. Clear definition combined with social service as sanction for hate speech for individuals’ effective sanction. Aggravating circumstances for hate speech of organisations.
12. EU law → a clear definition on hate speech with harmonise sanction and an obligation for European countries to comply.
13. One law = Reduce + simplify.
14. Prohibition of incitement of hatred in the employment context: Harassment of workers due to certain grounds (sexism, racism, homophobia) or declarations of not hiring individuals belonging to certain groups, or discouraging them from application to certain positions.
15. I would encourage positive speech.
16. Emigration + racism.
17. To criminalize false news making, from entrance right parties / political reasons.
18. A law fully transposing the framework decision on racism, to all found in Article 2A of the EU Charter for Fundamental Rights.
19. One European law to fight against a common approach of hate repression.
20. Facebook should allow LE and HATE get all requested data.

The answers above can be categorized in more than one ways. One such way is the following:

A. Freedom of speech: **4** (*20%*)
B. Clear/simple definition of hate speech: **4** (*20%*)
C. No need for more laws: **3** (*15%*)
D. Internet industry responsibilities: **3** (*15%*)
E. Racial discrimination: **2** (*10%*)
F. Legislation about religion: **1** (*5%*)
G. Sexual discrimination: **1** (*5%*)
H. Workplace discriminations: **1** (5%)
I. Spreading 'false news': **1** (*5%*)

4.5.4 Panel I / Question 4

Question: “*Are the reporting mechanisms of hate speech today enough? If not, how would you improve them?*”.

MANDOLA
Monitoring and Detecting Online Hate Speech

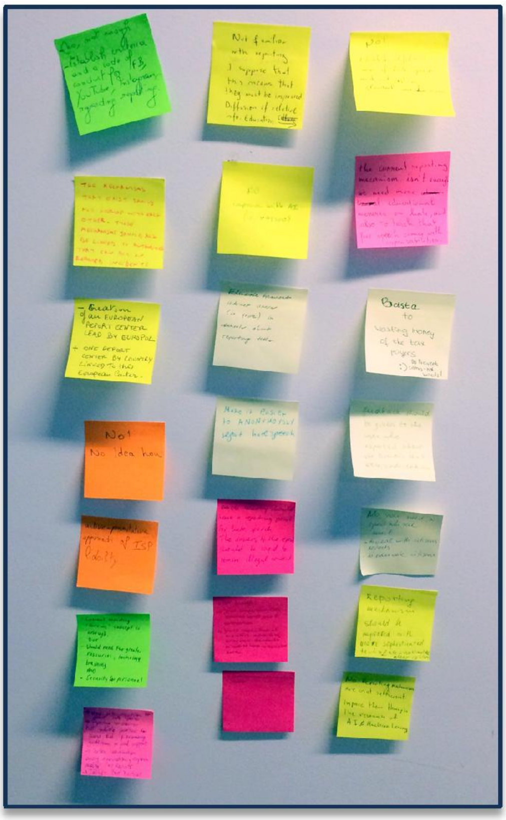
Brainstorming Panel I

- Are the reporting mechanisms of hate speech today enough?
  - If not, how would you improve them?

mandola-contact@mandola-project.eu
www.mandola-project.eu
22

Answers:

1. No, not enough. Establish and a code of conduct for FB, YouTube, Instagram, regarding reporting.
2. Not familiar with reporting. I suppose that this means that they must be improved. Diffusion of relative info. Education.
3. NO! Establish different ??????[5] of hate speech and not ???????? ???????????[5].
4. The mechanisms that exist should not overlap with any other. These mechanisms should also be limited to authorities they can act on referred incidents.
5. No improvement with AI.
6. The current reporting mechanism isn’t enough. We need more educational measures on hate and also to teach that free speech comes with responsibilities.
7. Creation of a European reporting Centre lead by Europol + one report centre by country linked to this European Centre.
8. Educate internet user (i.e. you) in school about reporting tool.
9. Basta to wasting money of the tax payers. Prevent crimes – not words! ☺
10. No! No idea how.
11. Make it easier to ANONYMOUSLY report hate speech.
12. Feedback should be given to the user who reported about the actions that were undertaken.
13. Active – preventive approach of ISP liability.
14. Each country should have a reporting point for hate speech. The answers to the report should be rapid to remove illegal content.
15. No, you need a specialized unit, to deal with citizens reports and to educate citizens.
16. Current reporting “HOTLINE”, concept is enough. BUT: Would need far greater resources, technology, training AND security for personnel.
17. Not enough: Joint cooperation CSOs, national Government and IT companies. Build capacities of monitors according [to] European standards so that to have comparable data. Involvement of law enforcement.
18. Reporting mechanism should be imposed with more sophisticated techniques – automatic alert systems.

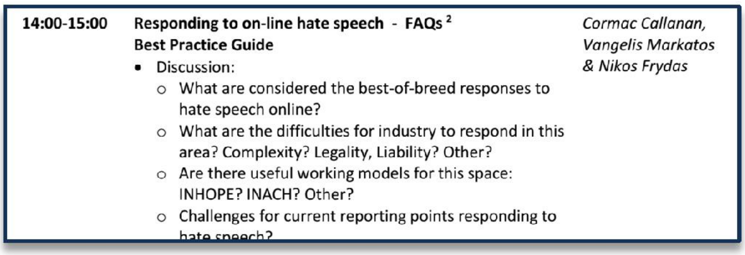


[5] Not clear

19. More automatization to detect hate speech. Reporting mechanism that provide feedback to users, and ……. Guide them to find support. Better coordination among organizations, companies daily w/ Reports → Identify Best Practices.
20. No. Reporting mechanisms are not sufficient. Improve them through the research of AI / Machine Learning.

4.6 Brainstorming Panel II

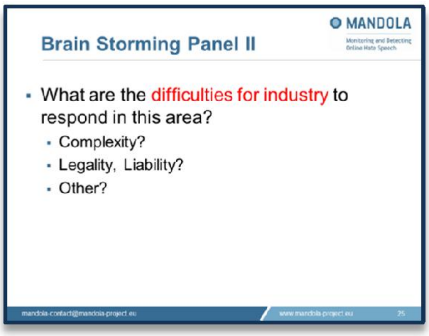
In this session, four questions were given to the AB. For each question, the members wrote their answers on sticky notes, which were then collected, read, displayed on the wall and recorder for processing.



14:00-15:00	**Responding to on-line hate speech - FAQs** [2] **Best Practice Guide** • Discussion: ○ What are considered the best-of-breed responses to hate speech online? ○ What are the difficulties for industry to respond in this area? Complexity? Legality, Liability? Other? ○ Are there useful working models for this space: INHOPE? INACH? Other? ○ Challenges for current reporting points responding to hate speech?	*Cormac Callanan, Vangelis Markatos & Nikos Frydas*

4.6.1 Panel II / Question 1

Question: "*What are the difficulties for industry to respond in this area? Complexity? Legality, Liability?, or Other?*".



Brain Storming Panel II

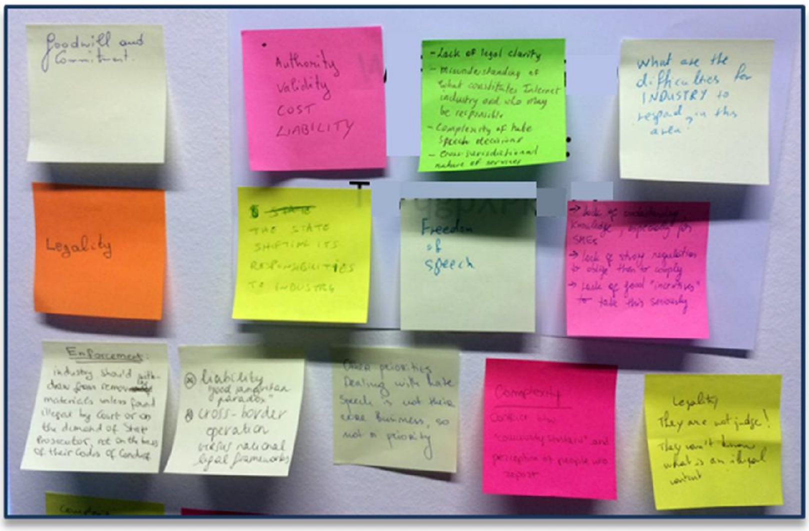
MANDOLA
Monitoring and Detecting Online Hate Speech

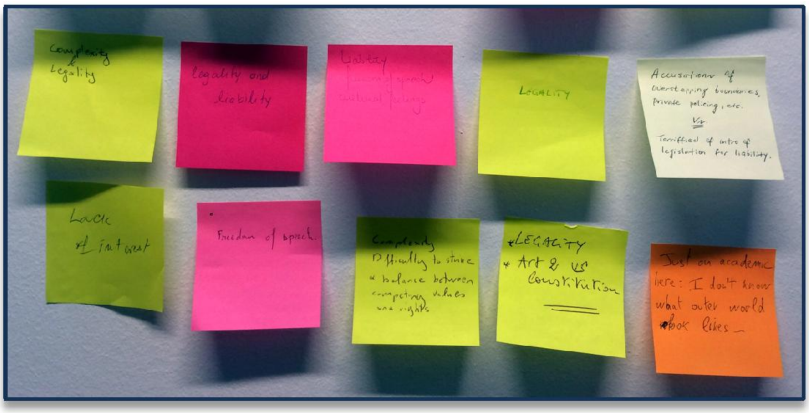
- What are the difficulties for industry to respond in this area?
  - Complexity?
  - Legality, Liability?
  - Other?

mandola-contact@mandola-project.eu www.mandola-project.eu 25

Answers:

1. Goodwill and Commitment.
2. Authority, Validity, Cost, Liability.
3. Lack of legal clarity, Misunderstanding of what constitutes Internet industry and who may be responsible, Complexity of hate speech decisions, Cross-jurisdictional nature of services.
4. Legality.
5. The state shifting its responsibility to industry.
6. Freedom of speech.
7. Lack of understanding, knowledge, especially for SMEs. Lack of strong regulation to oblige them to comply. Lack of good "incentives" to take this seriously.
8. Enforcement: Industry should withdraw from removing material unless found illegal by court or on the demand of state prosecutors, not on the basis of their codes of conduct.
9. Liability "good Samaritan paradox". Cross-border operation versus national legal frameworks.
10. Other priorities dealing with hate speech is not their core business, so not a priority.
11. Complexity: Conflict btw "community standard" and perception of people who report.
12. Legality: They are not judge! They can't know what an illegal content is.

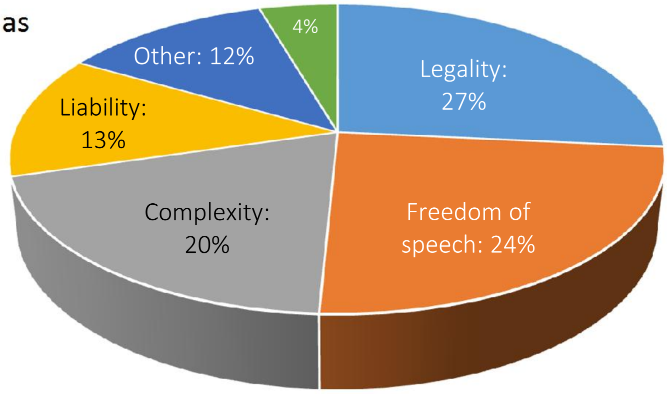

Goodwill and Commitment.
Authority
Validity
COST
LIABILITY
- Lack of legal clarity
- Misunderstanding of what constitutes Internet industry and who may be responsible
- Complexity of hate speech decisions
- Cross-jurisdictional nature of services
What are the difficulties for INDUSTRY to respond in this area?
Legality
THE STATE SHIFTING ITS RESPONSIBILITIES TO INDUSTRY
Freedom of speech
→ lack of strong regulation to oblige them to comply
→ lack of good "incentives" to take this seriously
Enforcement:
industry should withdraw from removing materials unless found illegal by Court or on the demand of State Prosecutor, not on the basis of their Codes of Conduct
liability "good samaritan paradox"
cross-border operation versus national legal frameworks
Other priorities Dealing with hate Speech is not their core Business, so not a priority
Complexity
Legality
They are not judge!
They can't know what is an illegal content


Complexity
Legality
Legality and liability
Legality
Accusations of overstepping boundaries, private policing, etc.
Terrified of intro of legislation for liability.
Lack of interest
Freedom of speech
Complexity
Difficulty to strike a balance between competing values and rights
LEGALITY
Art 2 US Constitution
Just an academic here: I don't know what outer world looks likes

13. Complexity. Legality.
14. Legality and liability.
15. Liability, freedom of speech, cultural feelings.
16. Legality.
17. Accusation of overstepping boundaries, private policing, etc. Terrified of intro of legislation for liability.
18. Lack of interest.
19. Freedom of speech.
20. Complexity. Difficulty to strike a balance between competing values and rights.
21. Legality. Article 2 US Constitution.
22. Just an academic here: I don't know what outer world looks like.

The responses above can be categorized as following:

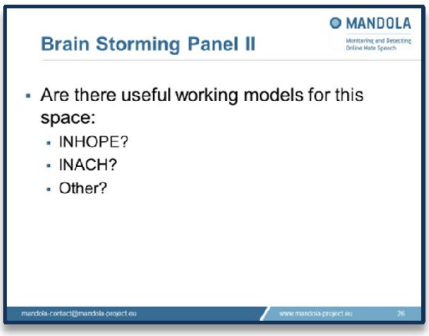
A. Legality: 1+1+1+ ½ + ½ + 1 + ⅓ + ½ = **5⅚** (*27%*)
B. Freedom of speech: ⅓ + 1 + 1 + 1 + 1 + 1 = **5⅓** (*24%*)
C. Complexity/Cost: 1 + ½ + 1 +⅓ + ⅓ + ½ + 1 = **4⅔** (*20%*)
D. Liability: ½ + 1 + 1 + ⅓ = **2⅚** (*13%*)
E. Other: ⅓ + 1 + 1 = **2⅓** (*12%*)
F. Don't know: **1** (*4%*)



4.6.2 Panel II / Question 2

Question: “*Are there useful working models for this space: INHOPE?, INACH?, Other?*”.

Answers:



Brain Storming Panel II

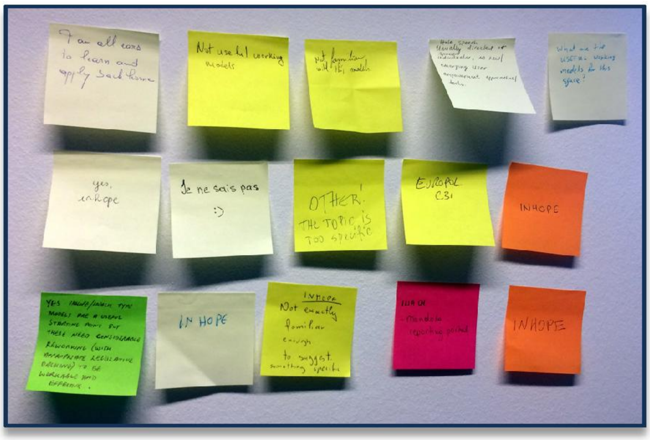
- Are there useful working models for this space:
  - INHOPE?
  - INACH?
  - Other?

mandola-contact@mandola-project.eu

www.mandola-project.eu

26

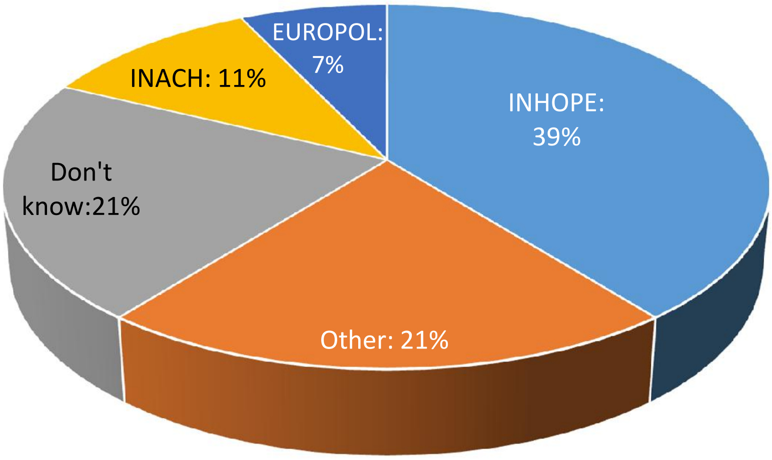
1. INHOPE: 1+1+1+1
2. I am all ears to learn and apply back home: 1
3. Not useful working models: 1
4. Not familiar with these models: 1
5. Hate speech usually directed at specific individuals, no new/emerging user empowerment approaches/tools: 1
6. Je ne sais pas ☺: 1
7. OTHER! The topic is too specific: 1
8. EUROPOL, C3I: 1
9. Yes INHOPE / INACH type models are a useful starting point, but these need considerable reworking (with appropriate legislative backing) to be workable and effective: 1
10. INHOPE. Not exactly familiar enough to suggest something specific.
11. INACH. Mandola reporting portal: 1



I am all ears to learn and apply back home
Not useful working models
Not familiar with these models
yes, inhope
Je ne sais pas :)
OTHER! THE TOPIC IS TOO SPECIFIC
EUROPOL C3I
INHOPE
IN HOPE
INHOPE Not exactly familiar enough to suggest something specific
INACH Mandola reporting portal
INHOPE

The responses above can be categorized as following:

A. INHOPE: 4 + ½ + 1 = **5½** (*39%*)
B. Other: 1 + 1 + 1 = **3** (*21%*)
C. Don’t know: 1 + 1 + 1 = **3** (*21%*)
D. INACH: ½ + 1 = **1½** (*11%*)
E. EUROPOL/C3I: **1** (*7%*)



4.6.3 Panel II / Question 3

Question: "*"What are the Challenges for current reporting points responding to hate speech?*".

MANDOLA Monitoring and Detecting Online Hate Speech

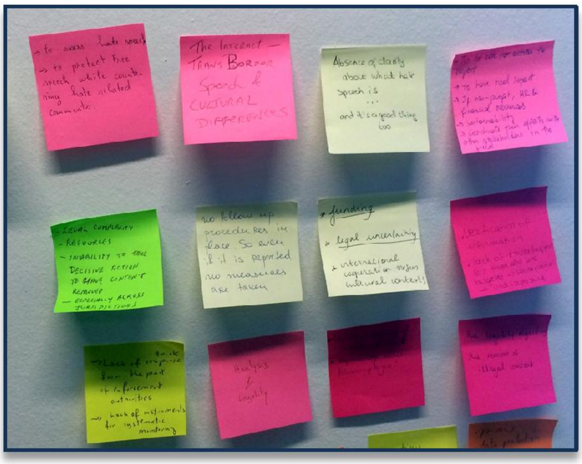
Brain Storming Panel II

- What are the Challenges for current reporting points responding to hate speech?

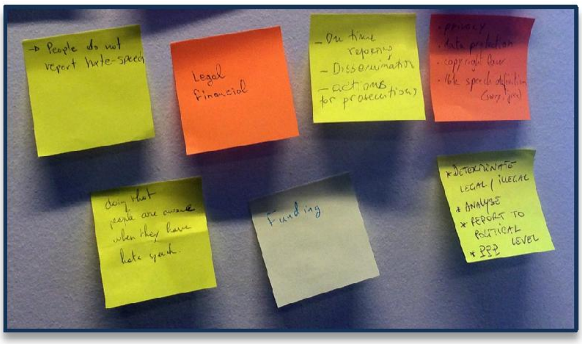
mandola-contact@mandola-project.eu www.mandola-project.eu 27

Answers:

1. To assess hate speech. To protect free speech while countering hate-related comments.
2. The Internet – Trans border. Speech & Cultural differences.
3. Absence of clarity about what hated speech is, and it's a good thing too.
4. To be able to action the report. To have real impact. If non-profit, HR & financial resources. Sustainability. Coordinate their efforts with other stakeholders in the field.
5. Legal complexity. Resources. Inability to take decisive action to have content removed. Especially across jurisdictions.
6. No follow up procedures in place. So even if it is reported no measures are taken.
7. Funding. Legal uncertainty. International cooperation versus cultural contexts.
8. Verification of information. Lack of service support for those who are targeted → consequences → under-reporting.


→ to assess hate speech
→ to protect free speech while countering hate related comments.
The Internet – TRANS BORDER Speech & CULTURAL DIFFERENCES
Absence of clarity about what hate speech is ... and it's a good thing too
→ To have real impact
→ If non-profit, HR & financial resources
→ Sustainability
→ Coordinate their efforts with other stakeholders in the field
– LEGAL COMPLEXITY
– RESOURCES
– INABILITY TO TAKE DECISIVE ACTION TO HAVE CONTENT REMOVED
– ESPECIALLY ACROSS JURISDICTIONS
no follow up procedures in place. So even if it is reported no measures are taken
* funding
* legal uncertainty
* international cooperation versus cultural contexts
→ Lack of quick response from the part of enforcement authorities
→ Lack of instruments for systematic monitoring
Analysis & Legality
illegal content

9. Lack of quick response from the part of enforcement authorities. Lack of instruments for systematic monitoring.
10. Analysis & Legality.
11. Legality or illegality of ??? ??? Responsibility what follow up to give?
12. The legality regarding the removal of illegal content.
13. People do not report hate speech.
14. Legal. Financial.
15. On time reporting. Dissemination. Actions for prosecutions.
16. Privacy. Data protection. Copyright law. Hate speech definition (sorry, joke).


→ People do not report hate-speech
Legal Financial
– On time reporting
– Dissemination
– actions for prosecutions
· privacy
· data protection
· copyright law
· Hate speech definition (sorry, joke)
doing what people are aware when they have hate speech.
Funding
* DETERMINATE LEGAL / ILLEGAL
* ANALYSE
* REPORT TO POLITICAL LEVEL
* PPP

17. Doing what people are aware when they have hate speech.
18. Funding.
19. Determine legal/illegal. Analyse. Report to political level. PPP.

The responses above can be categorized as following:

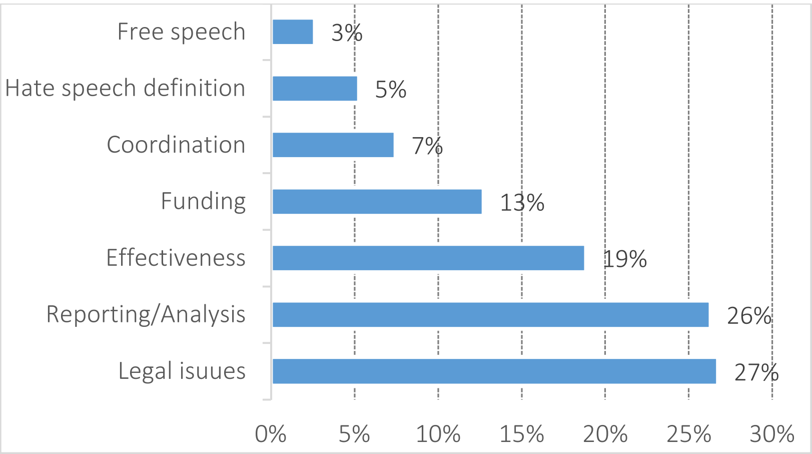
A. Legal issues: ½ + ¼ + ⅓ + ½ + ½ + 1 + ½ + 1 + ½ = 5(1/12)
B. Reporting/Analysis: ½ + 1 + 1+ ½ + ½ + 1 + ½ = 5
C. Effectiveness: ⅓ + ¼ + 1 + ½ + ½ + 1 = 3(7/12)

D. Funding: ⅓ + 1 + ½ + ⅓ + ¼ = 2(5/12)

E. Coordination: ½ + ⅓ + ¼ + ⅓ = 1(5/12)

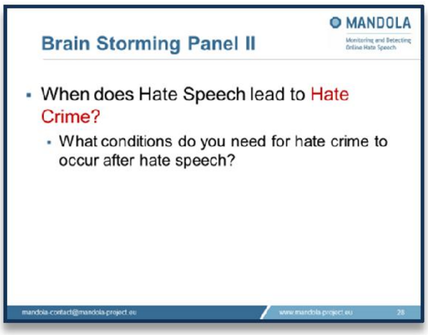
F. Hate speech definition: 1

G. Free speech: ½



4.6.4 Panel II / Question 4

Question: “*When does Hate Speech lead to Hate Crime? What conditions to you need for hate crime to occur after hate speech?*”.



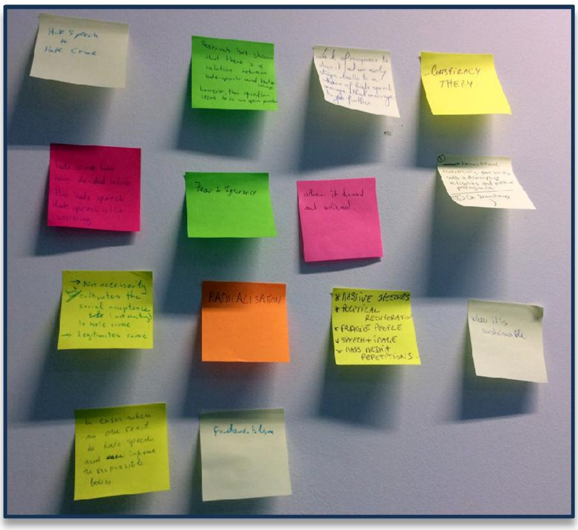
Brain Storming Panel II

- When does Hate Speech lead to Hate Crime?
  - What conditions do you need for hate crime to occur after hate speech?

mandola-contact@mandola-project.eu www.mandola-project.eu 28

Answers:

1. Research has shown that there is a relation between hate-speech and hate-crime however, this question seems to be an open problem.
2. Lack of response to stop it at an early stage leads to a chain of hate speech message that encourage to go further.
3. Conspiracy theory.
4. Hate crime has been decided before the hate speech. Hate speech is the warning.
5. Fear & Ignorance.
6. When it heard and believed.
7. Transitional economically poor societies with a domineering religious and political propaganda.
8. Not necessarily but cultivates the social acceptance (not reacting) to hate crime. Legitimates crime.
9. Radicalization.
10. Passive speeches. Political recuperation. Fragile people. Speech + image. Mass media repetitions.
11. When it is sustainable.
12. In cases when no one reacts to hate speech and informs the responsible bodies.
13. Fundamentalism.



5 Conclusions & Lessons Learned

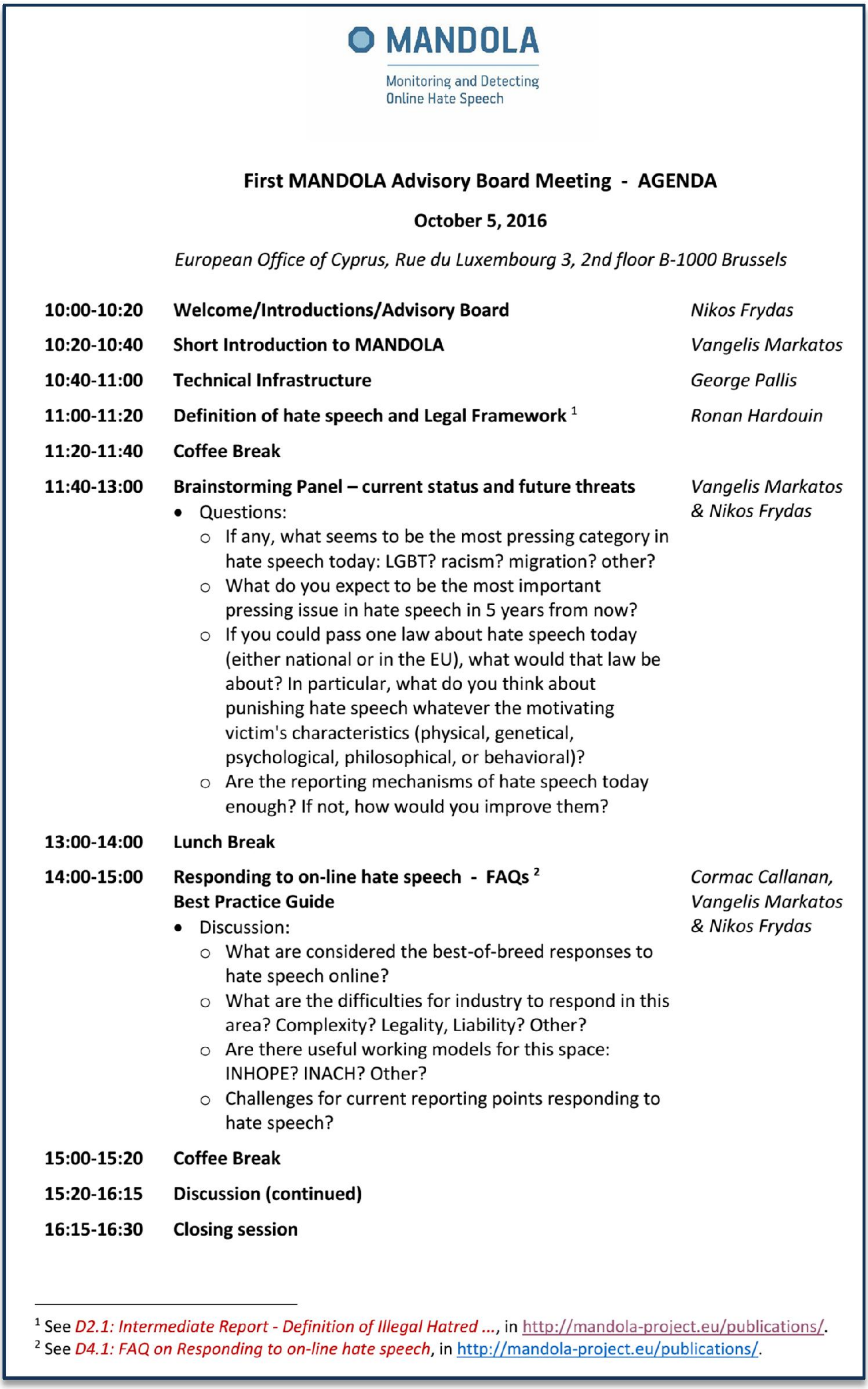
This chapter gives the conclusions and lessons learned from AB1.

The size of the AB (16 external and 9 internal members) appears to be working very well, as the time available (6.5 hours gross time) allowed each member to be able to contribute more than one times. If space allows, AB2 may grow to 20+9 members.

Some topics, if presented adequately, may benefit from an AB debate, but the time available would allow only one (perhaps two) such cases.

The 'sticky-notes' brainstorming sessions are very productive and allow for the collection of hard evidence from each member. In AB1 it was possible to conduct eight such sessions. The results are available in the current report (§4.5 in p. 21 and §4.6 in p. 28).

6 Appendix A: Agenda of AB1 (Advisory Board Meeting 1)



MANDOLA

Monitoring and Detecting
Online Hate Speech

First MANDOLA Advisory Board Meeting - AGENDA

October 5, 2016

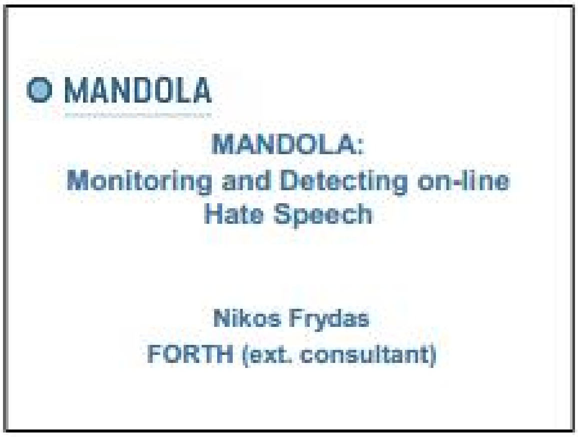
European Office of Cyprus, Rue du Luxembourg 3, 2nd floor B-1000 Brussels

10:00-10:20	**Welcome/Introductions/Advisory Board**	*Nikos Frydas*
10:20-10:40	**Short Introduction to MANDOLA**	*Vangelis Markatos*
10:40-11:00	**Technical Infrastructure**	*George Pallis*
11:00-11:20	**Definition of hate speech and Legal Framework** [1]	*Ronan Hardouin*
11:20-11:40	**Coffee Break**	
11:40-13:00	**Brainstorming Panel – current status and future threats** • Questions: ○ If any, what seems to be the most pressing category in hate speech today: LGBT? racism? migration? other? ○ What do you expect to be the most important pressing issue in hate speech in 5 years from now? ○ If you could pass one law about hate speech today (either national or in the EU), what would that law be about? In particular, what do you think about punishing hate speech whatever the motivating victim's characteristics (physical, genetical, psychological, philosophical, or behavioral)? ○ Are the reporting mechanisms of hate speech today enough? If not, how would you improve them?	*Vangelis Markatos & Nikos Frydas*
13:00-14:00	**Lunch Break**	
14:00-15:00	**Responding to on-line hate speech - FAQs** [2] **Best Practice Guide** • Discussion: ○ What are considered the best-of-breed responses to hate speech online? ○ What are the difficulties for industry to respond in this area? Complexity? Legality, Liability? Other? ○ Are there useful working models for this space: INHOPE? INACH? Other? ○ Challenges for current reporting points responding to hate speech?	*Cormac Callanan, Vangelis Markatos & Nikos Frydas*
15:00-15:20	**Coffee Break**	
15:20-16:15	**Discussion (continued)**	
16:15-16:30	**Closing session**	

[1] See *D2.1: Intermediate Report - Definition of Illegal Hatred ...*, in http://mandola-project.eu/publications/.

[2] See *D4.1: FAQ on Responding to on-line hate speech*, in http://mandola-project.eu/publications/.

7 Appendix B: Advisory Board presentation



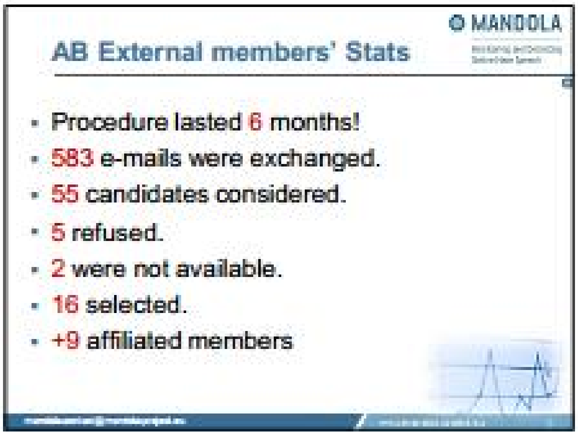
MANDOLA

MANDOLA:
Monitoring and Detecting on-line
Hate Speech

Nikos Frydas
FORTH (ext. consultant)

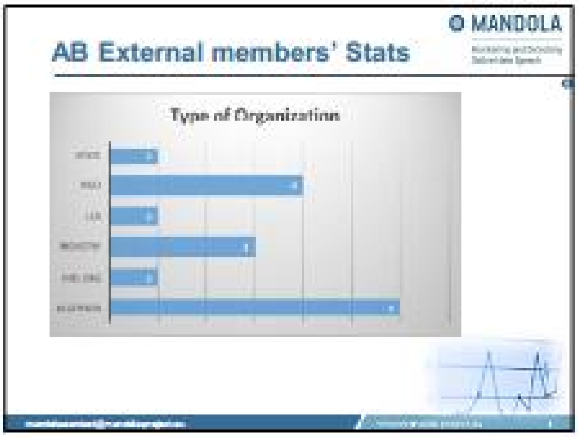
Welcome!

- This is the 1st Advisory Board Meeting.
- Welcome!
- 2nd AB meeting co-ords: Brussels ~6/2017

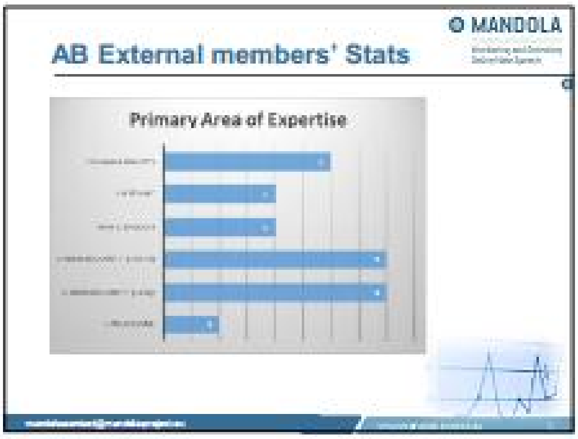


AB External members' Stats

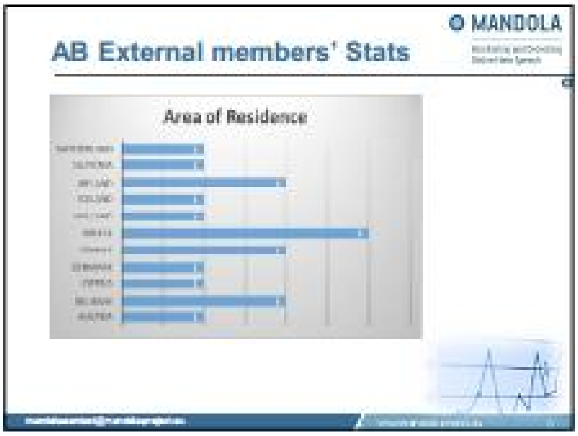
- Procedure lasted 6 months!
- 583 e-mails were exchanged.
- 55 candidates considered.
- 5 refused.
- 2 were not available.
- 16 selected.
- +9 affiliated members



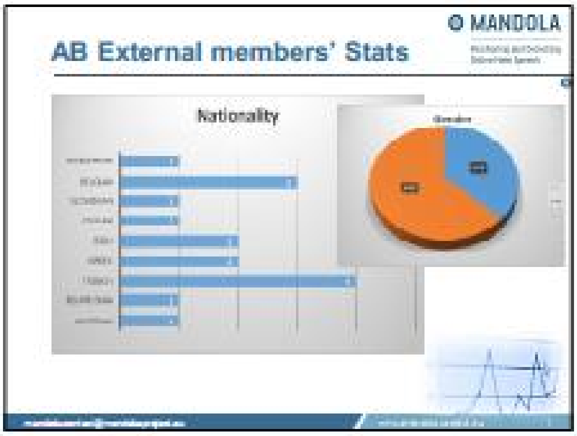
AB External members' Stats



AB External members' Stats



AB External members' Stats

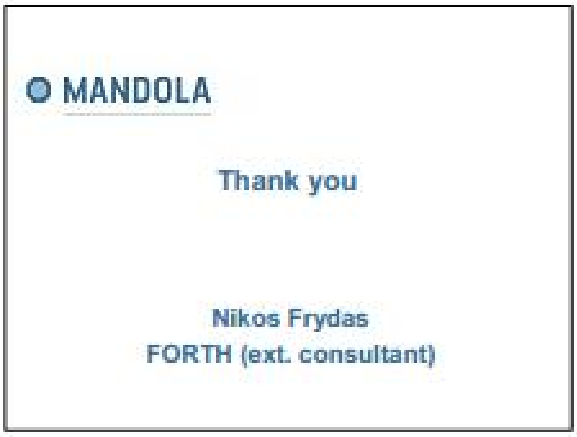


AB External members' Stats

Welcome!

- Let us now move on with our Agenda.
- Before that, may we introduce each other?

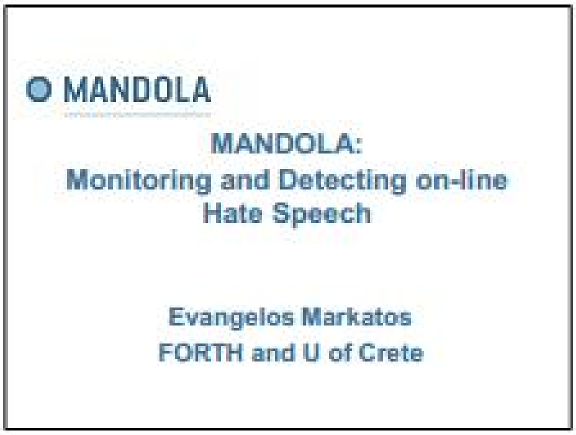
time to start



Thank you

Nikos Frydas
FORTH (ext. consultant)

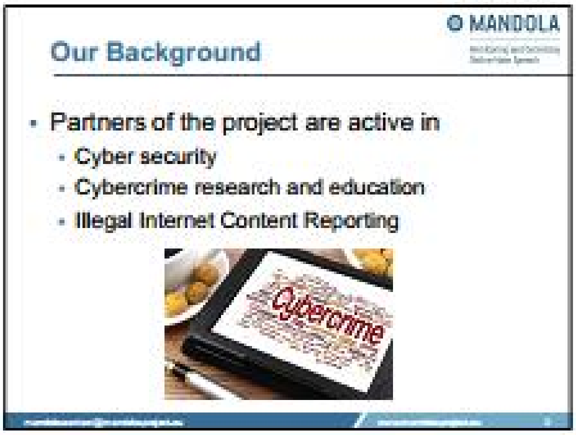
8 Appendix C: Introduction to MANDOLA presentation



MANDOLA

MANDOLA:
Monitoring and Detecting on-line
Hate Speech

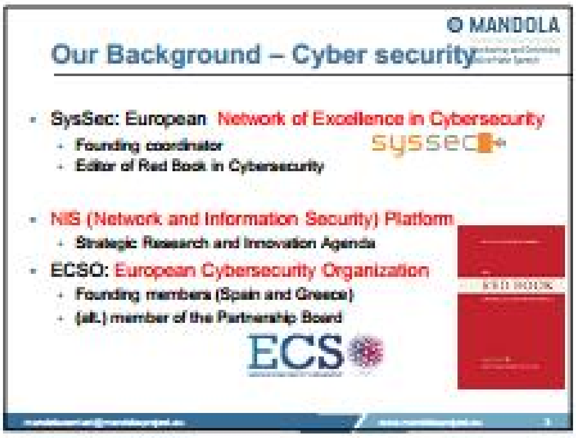
Evangelos Markatos
FORTH and U of Crete



Our Background

- Partners of the project are active in
  - Cyber security
  - Cybercrime research and education
  - Illegal Internet Content Reporting

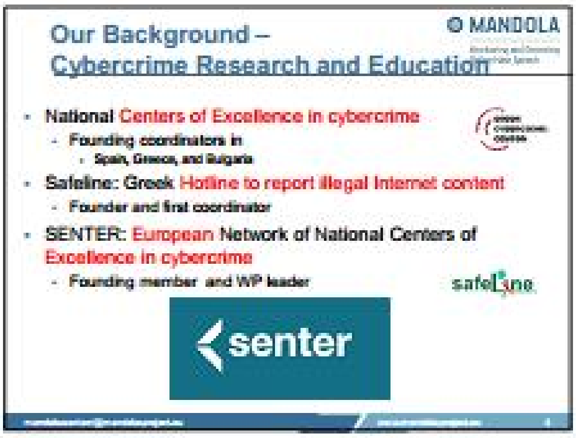
Cybercrime



Our Background – Cyber security

- SysSec: European Network of Excellence in Cybersecurity
  - Founding coordinator
  - Editor of Red Book in Cybersecurity
- NIS (Network and Information Security) Platform
  - Strategic Research and Innovation Agenda
- ECSO: European Cybersecurity Organization
  - Founding members (Spain and Greece)
  - (alt.) member of the Partnership Board

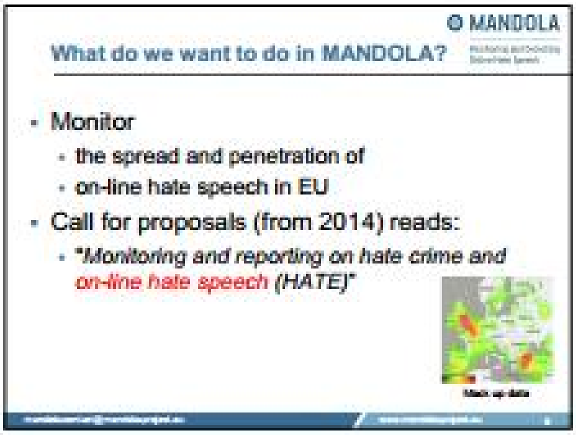
syssec
RED BOOK
ECS



Our Background – Cybercrime Research and Education

- National Centers of Excellence in cybercrime
  - Founding coordinators in
    - Spain, Greece, and Bulgaria
- Safeline: Greek Hotline to report illegal Internet content
  - Founder and first coordinator
- SENTER: European Network of National Centers of Excellence in cybercrime
  - Founding member and WP leader

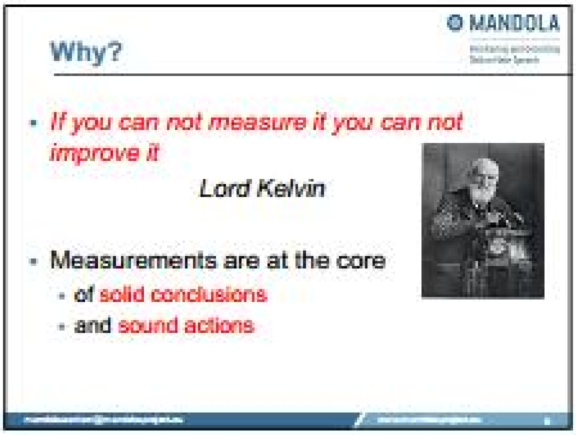
safeline
senter



What do we want to do in MANDOLA?

- Monitor
  - the spread and penetration of
  - on-line hate speech in EU
- Call for proposals (from 2014) reads:
  - *"Monitoring and reporting on hate crime and on-line hate speech (HATE)"*

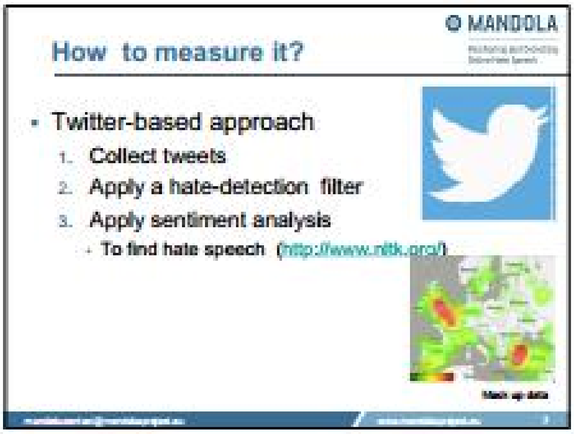
Mock up data



Why?

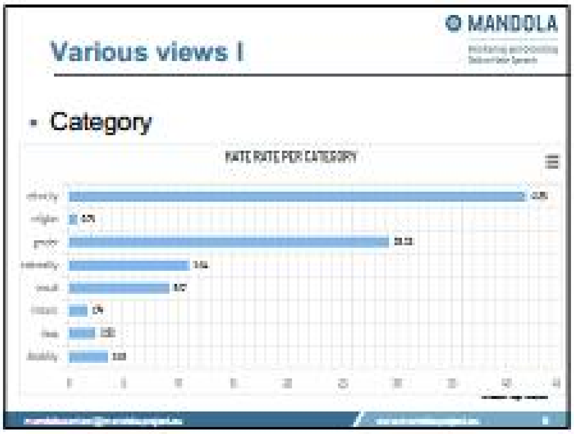
- *If you can not measure it you can not improve it*

  Lord Kelvin
- Measurements are at the core
  - of solid conclusions
  - and sound actions



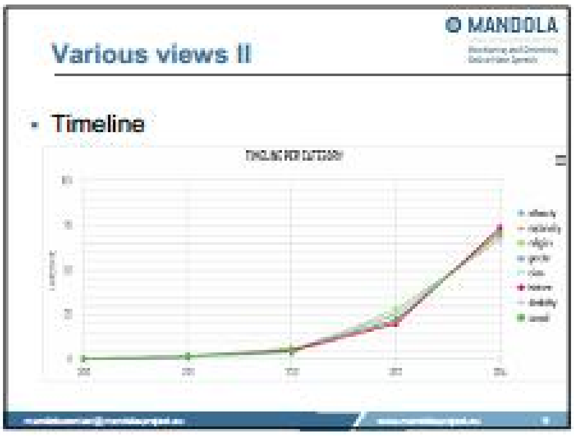
How to measure it?

- Twitter-based approach
  1. Collect tweets
  2. Apply a hate-detection filter
  3. Apply sentiment analysis
     - To find hate speech (http://www.nltk.org/)



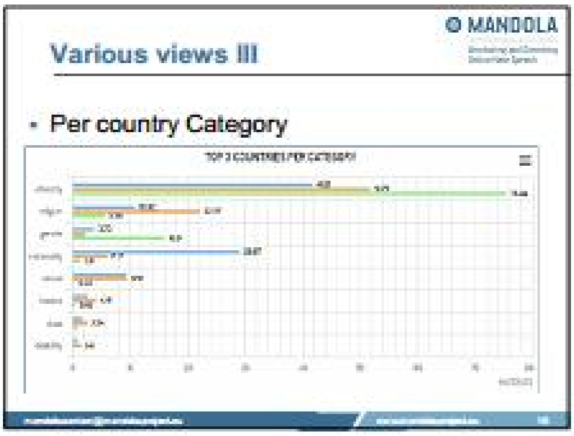
Various views I

- Category



Various views II

- Timeline



Various views III

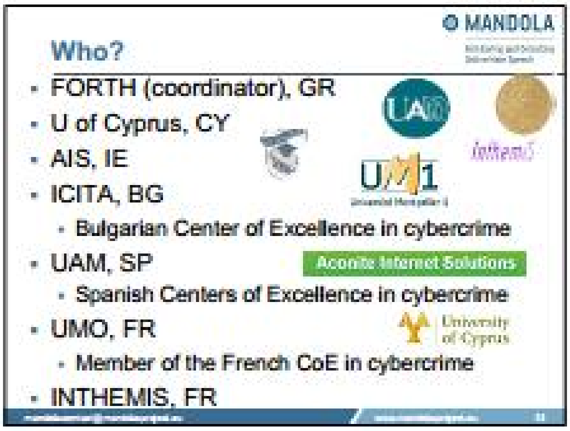
- Per country Category

Other activities? Frequently Asked Questions

- FAQ Book
- What to do?
  - *I believe I have encountered hate speech. What can I do about it?*
  - *Why words can be harmful and sometimes dangerous?*
  - *I believe I have been a target of hate speech? What are my rights as a victim?*
- More at the talk at 14:00

Other activities? Legal

- Definition
  - *What is hate speech*
- Legal Framework
  - *In Member States*
- More at the talk at 11:00



Who?

- FORTH (coordinator), GR
- U of Cyprus, CY
- AIS, IE
- ICITA, BG
  - Bulgarian Center of Excellence in cybercrime
- UAM, SP
  - Spanish Centers of Excellence in cybercrime
- UMO, FR
  - Member of the French CoE in cybercrime
- INTHEMIS, FR

MANDOLA
Aconite Internet Solutions
University of Cyprus
inthemis
UM1

Why are we here today?

- We want your advice
  - Advisory Board
- You know a lot about this area
  - Can you share some of your knowledge?
  - Can you share some of your experience?
  - Some of your wisdom?

How are we going to get this advice?

- Interactive Brain storming sessions
- Post it notes
- Share your advice
  - Even half-baked ideas
  - All ideas!
    - In research all ideas are welcome

Brain storming sessions

- We are going to ask you questions:
  - e.g. "If you could pass one law in hate speech, what would it be?"
  - Think "out of the box"
  - Share your knowledge
  - Share your ideas

A final note

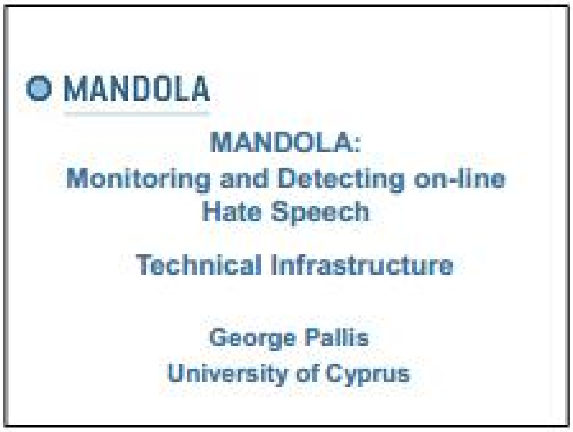
- If you want to make a short presentation
  - Of your activities
  - Let me know
  - We have some slots
  - < 5 minutes long…

MANDOLA

Thank you

Evangelos Markatos
FORTH and U of Crete

9 Appendix D: Technical Infrastructure presentation

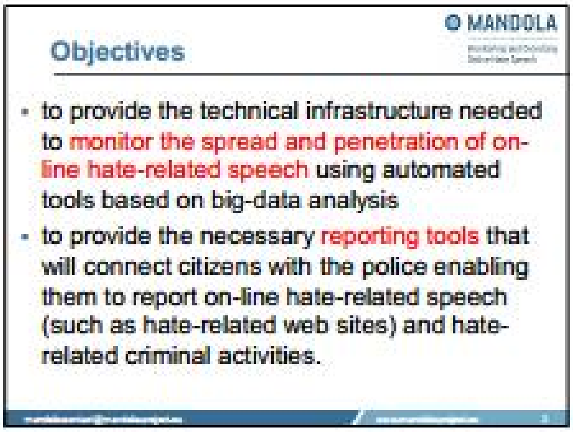


MANDOLA

MANDOLA:
Monitoring and Detecting on-line Hate Speech

Technical Infrastructure

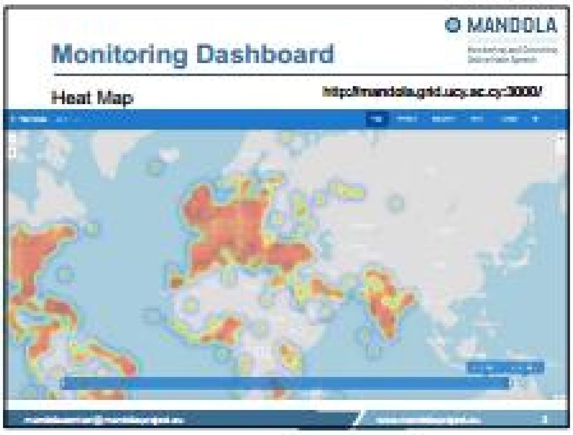
George Pallis
University of Cyprus



MANDOLA

Objectives

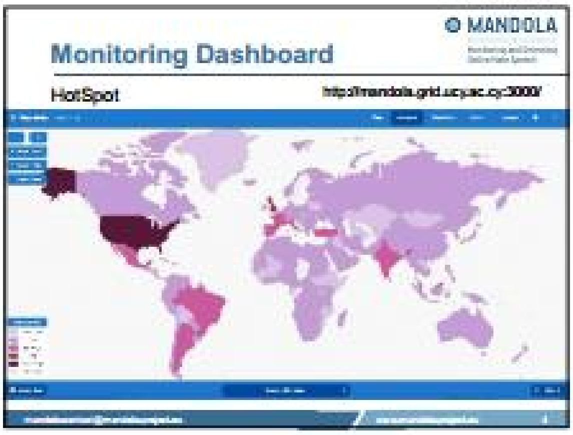
- to provide the technical infrastructure needed to monitor the spread and penetration of on-line hate-related speech using automated tools based on big-data analysis
- to provide the necessary reporting tools that will connect citizens with the police enabling them to report on-line hate-related speech (such as hate-related web sites) and hate-related criminal activities.



MANDOLA

Monitoring Dashboard

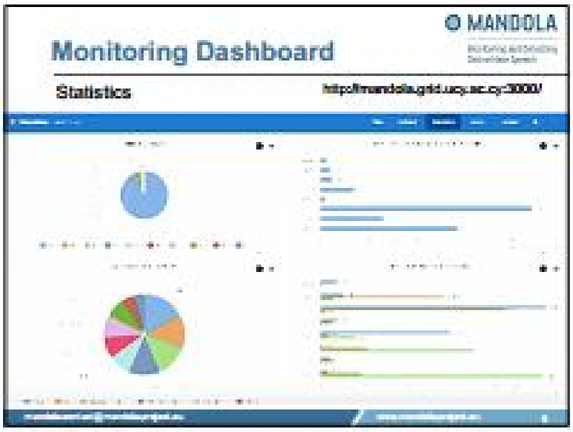
Heat Map http://mandola.grid.ucy.ac.cy:3000/



MANDOLA

Monitoring Dashboard

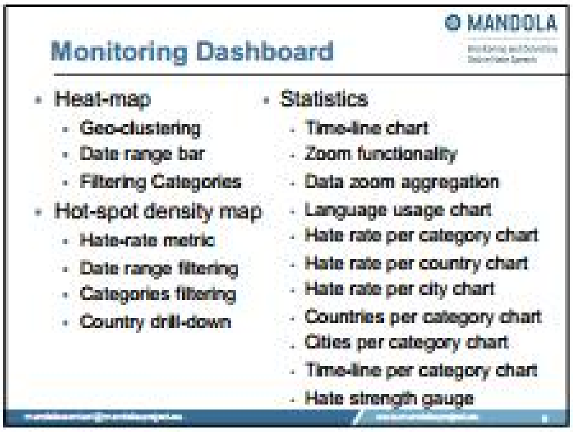
HotSpot http://mandola.grid.ucy.ac.cy:3000/



MANDOLA

Monitoring Dashboard

Statistics http://mandola.grid.ucy.ac.cy:3000/



MANDOLA

Monitoring Dashboard

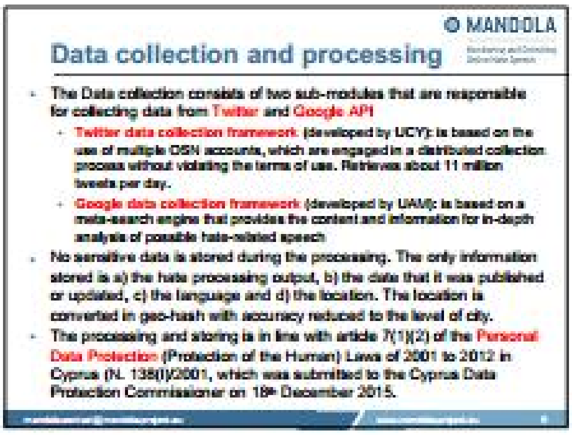
- Heat-map
  - Geo-clustering
  - Date range bar
  - Filtering Categories
- Hot-spot density map
  - Hate-rate metric
  - Date range filtering
  - Categories filtering
  - Country drill-down
- Statistics
  - Time-line chart
  - Zoom functionality
  - Data zoom aggregation
  - Language usage chart
  - Hate rate per category chart
  - Hate rate per country chart
  - Hate rate per city chart
  - Countries per category chart
  - Cities per category chart
  - Time-line per category chart
  - Hate strength gauge

Reporting Portal

http://mandola-project.eu/portal/

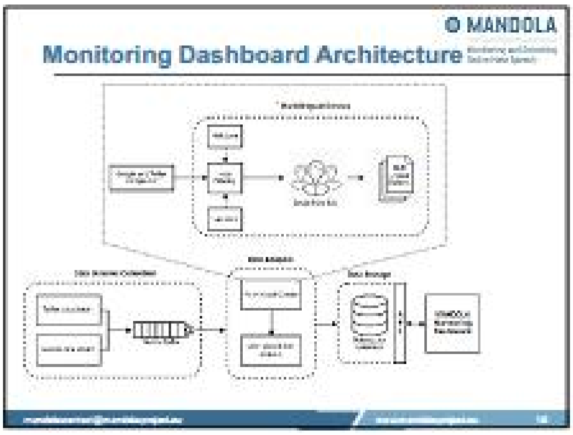
Reporting Portal

http://mandola-project.eu/portal/

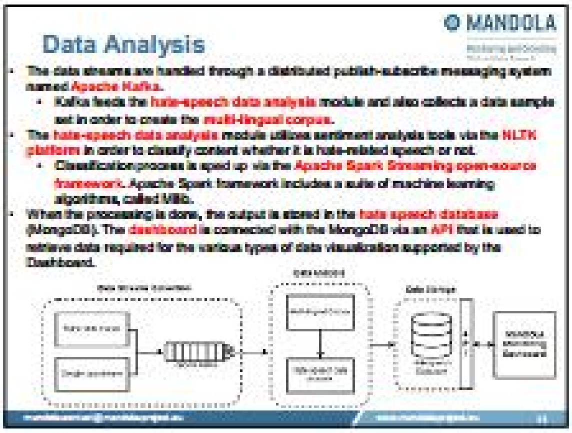


Data collection and processing

- The Data collection consists of two sub-modules that are responsible for collecting data from Twitter and Google API
  - Twitter data collection framework (developed by UCY): is based on the use of multiple OSN accounts, which are engaged in a distributed collection process without violating the terms of use. Retrieves about 11 million tweets per day.
  - Google data collection framework (developed by UAM): is based on a meta-search engine that provides the content and information for in-depth analysis of possible hate-related speech
- 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.
- 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, which was submitted to the Cyprus Data Protection Commissioner on 18th December 2015.



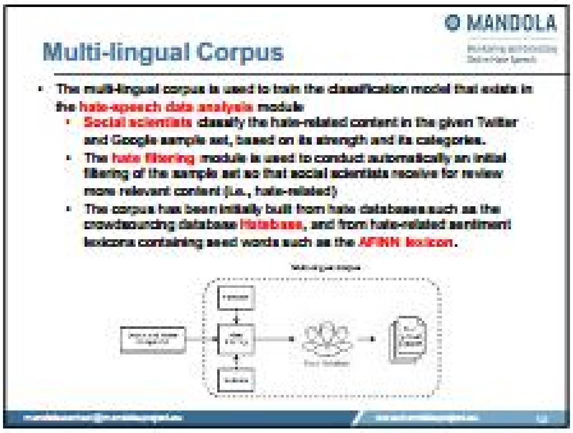
Monitoring Dashboard Architecture



Data Analysis

- The data streams are handled through a distributed publish-subscribe messaging system named Apache Kafka.
  - Kafka feeds the hate-speech data analysis module and also collects a data sample set in order to create the multi-lingual corpus.
- The hate-speech data analysis module utilizes sentiment analysis tools via the NLTK platform in order to classify content whether it is hate-related speech or not.
  - Classification process is sped up via the Apache Spark Streaming open-source framework. Apache Spark framework includes a suite of machine learning algorithms, called Mllib.
- When the processing is done, the output is stored in the hate speech database (MongoDB). The dashboard is connected with the MongoDB via an API that is used to retrieve data required for the various types of data visualization supported by the Dashboard.

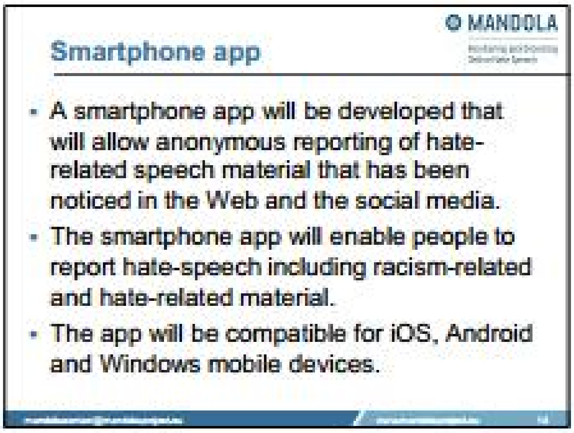
Data Streams Collection
Data Analysis
Data Storage
MANDOLA Monitoring Dashboard



Multi-lingual Corpus

- The multi-lingual corpus is used to train the classification model that exists in the hate-speech data analysis module
  - Social scientists classify the hate-related content in the given Twitter and Google sample set, based on its strength and its categories.
  - The hate filtering module is used to conduct automatically an initial filtering of the sample set so that social scientists receive for review more relevant content (i.e., hate-related)
  - The corpus has been initially built from hate databases such as the crowdsourcing database Hatebase, and from hate-related sentiment lexicons containing seed words such as the AFINN lexicon.

Social Scientists



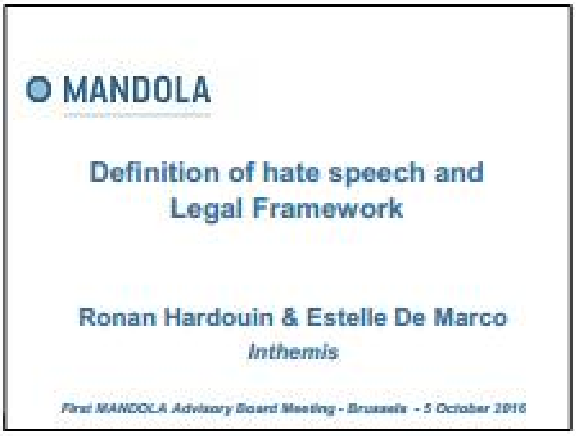
Smartphone app

- A smartphone app will be developed that will allow anonymous reporting of hate-related speech material that has been noticed in the Web and the social media.
- The smartphone app will enable people to report hate-speech including racism-related and hate-related material.
- The app will be compatible for iOS, Android and Windows mobile devices.

Thank you!

- Questions?

10 Appendix E: Definition of Hate Speech & Legal Framework presentation



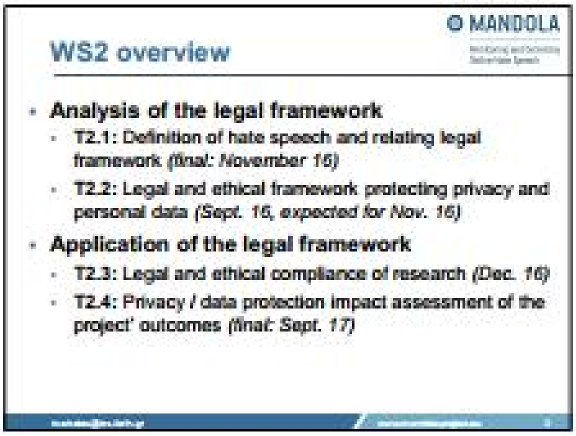
MANDOLA

Definition of hate speech and Legal Framework

Ronan Hardouin & Estelle De Marco

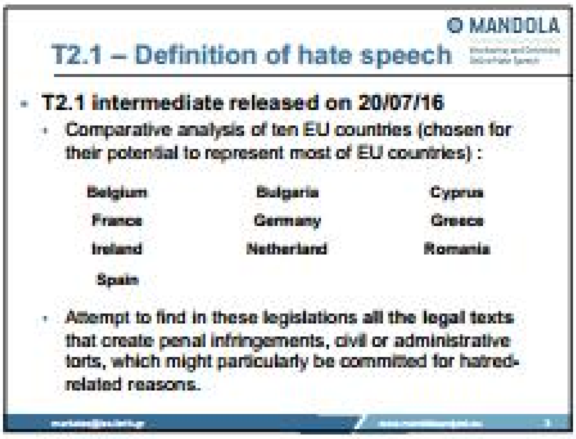
Inthemis

First MANDOLA Advisory Board Meeting - Brussels - 5 October 2016



WS2 overview

- Analysis of the legal framework
  - T2.1: Definition of hate speech and relating legal framework *(final: November 16)*
  - T2.2: Legal and ethical framework protecting privacy and personal data *(Sept. 16, expected for Nov. 16)*
- Application of the legal framework
  - T2.3: Legal and ethical compliance of research *(Dec. 16)*
  - T2.4: Privacy / data protection impact assessment of the project' outcomes *(final: Sept. 17)*

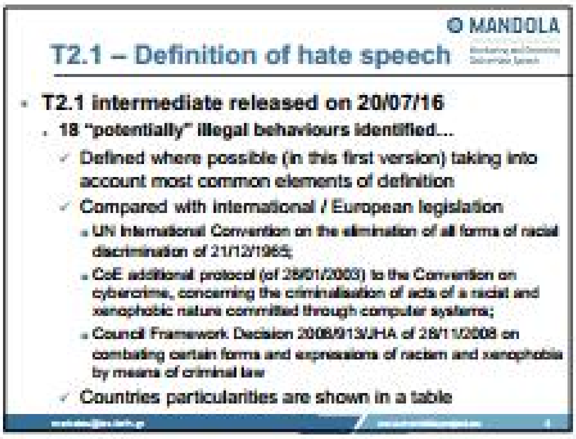


T2.1 – Definition of hate speech

- T2.1 intermediate released on 20/07/16
  - Comparative analysis of ten EU countries (chosen for their potential to represent most of EU countries) :

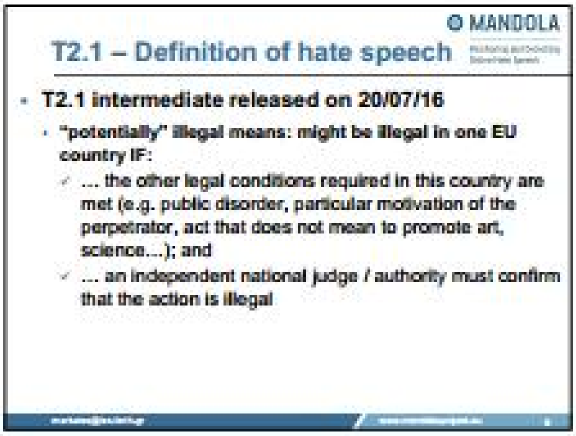
    Belgium, Bulgaria, Cyprus
    France, Germany, Greece
    Ireland, Netherland, Romania
    Spain

  - Attempt to find in these legislations all the legal texts that create penal infringements, civil or administrative torts, which might particularly be committed for hatred-related reasons.



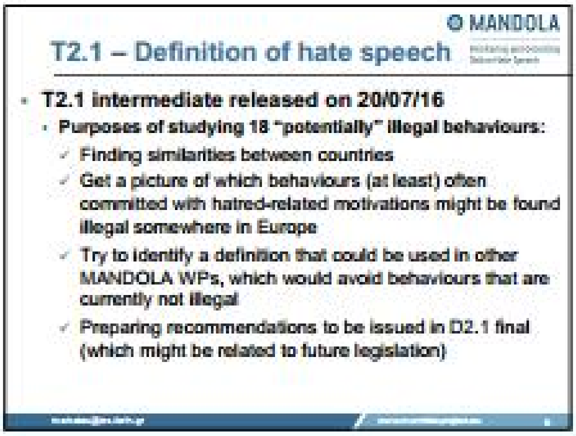
T2.1 – Definition of hate speech

- T2.1 intermediate released on 20/07/16
  - 18 "potentially" illegal behaviours identified...
    - ✓ Defined where possible (in this first version) taking into account most common elements of definition
    - ✓ Compared with international / European legislation
      - UN International Convention on the elimination of all forms of racial discrimination of 21/12/1965;
      - CoE additional protocol (of 28/01/2003) to the Convention on cybercrime, concerning the criminalisation of acts of a racist and xenophobic nature committed through computer systems;
      - Council Framework Decision 2008/913/JHA of 28/11/2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law
    - ✓ Countries particularities are shown in a table



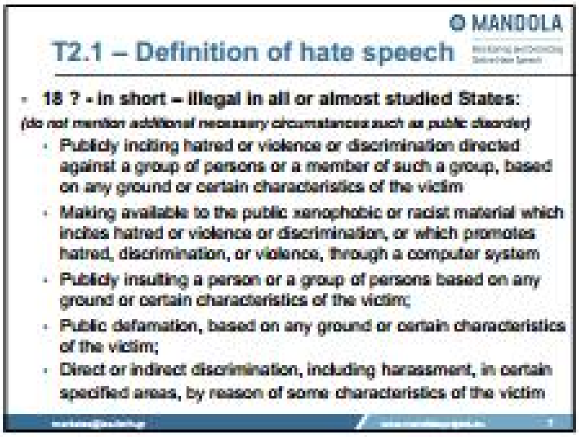
T2.1 – Definition of hate speech

- T2.1 intermediate released on 20/07/16
  - "potentially" illegal means: might be illegal in one EU country IF:
    - ✓ ... the other legal conditions required in this country are met (e.g. public disorder, particular motivation of the perpetrator, act that does not mean to promote art, science...); and
    - ✓ ... an independent national judge / authority must confirm that the action is illegal



T2.1 – Definition of hate speech

- T2.1 intermediate released on 20/07/16
  - Purposes of studying 18 "potentially" illegal behaviours:
    - ✓ Finding similarities between countries
    - ✓ Get a picture of which behaviours (at least) often committed with hatred-related motivations might be found illegal somewhere in Europe
    - ✓ Try to identify a definition that could be used in other MANDOLA WPs, which would avoid behaviours that are currently not illegal
    - ✓ Preparing recommendations to be issued in D2.1 final (which might be related to future legislation)

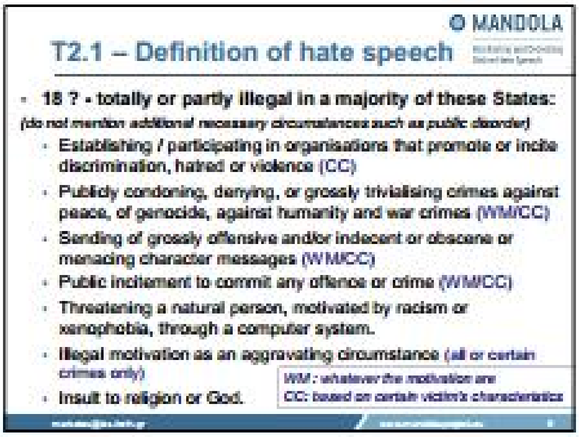


T2.1 – Definition of hate speech

- **18 ? - in short – illegal in all or almost studied States:**

(do not mention additional necessary circumstances such as public disorder)

  - Publicly inciting hatred or violence or discrimination directed against a group of persons or a member of such a group, based on any ground or certain characteristics of the victim
  - Making available to the public xenophobic or racist material which incites hatred or violence or discrimination, or which promotes hatred, discrimination, or violence, through a computer system
  - Publicly insulting a person or a group of persons based on any ground or certain characteristics of the victim;
  - Public defamation, based on any ground or certain characteristics of the victim;
  - Direct or indirect discrimination, including harassment, in certain specified areas, by reason of some characteristics of the victim



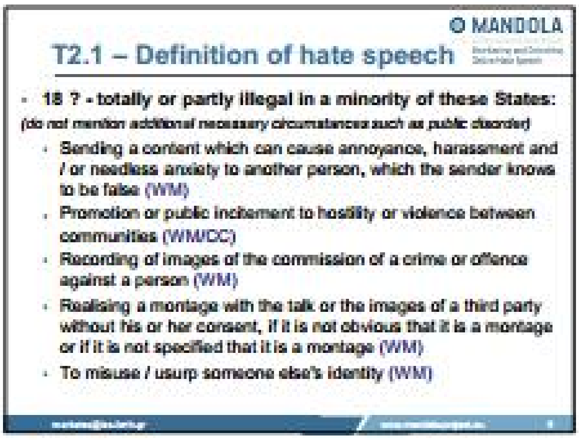
T2.1 – Definition of hate speech

- **18 ? - totally or partly illegal in a majority of these States:**

(do not mention additional necessary circumstances such as public disorder)

  - Establishing / participating in organisations that promote or incite discrimination, hatred or violence (CC)
  - Publicly condoning, denying, or grossly trivialising crimes against peace, of genocide, against humanity and war crimes (WM/CC)
  - Sending of grossly offensive and/or indecent or obscene or menacing character messages (WM/CC)
  - Public incitement to commit any offence or crime (WM/CC)
  - Threatening a natural person, motivated by racism or xenophobia, through a computer system.
  - Illegal motivation as an aggravating circumstance (all or certain crimes only)
  - Insult to religion or God.

WM : whatever the motivation are
CC: based on certain victim's characteristics

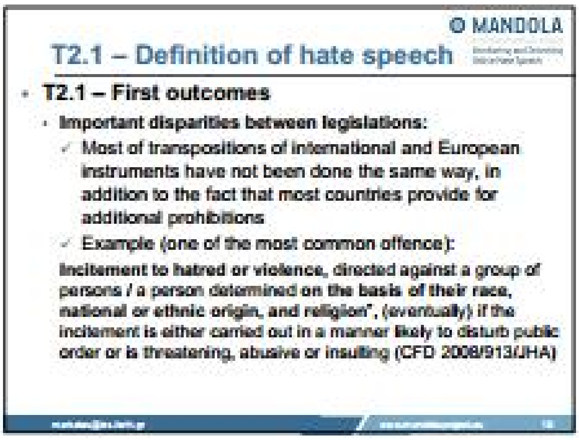


T2.1 – Definition of hate speech

- **18 ? - totally or partly illegal in a minority of these States:**

(do not mention additional necessary circumstances such as public disorder)

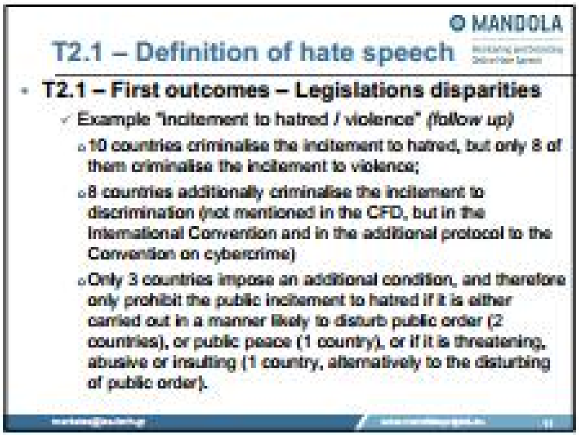
  - Sending a content which can cause annoyance, harassment and / or needless anxiety to another person, which the sender knows to be false (WM)
  - Promotion or public incitement to hostility or violence between communities (WM/CC)
  - Recording of images of the commission of a crime or offence against a person (WM)
  - Realising a montage with the talk or the images of a third party without his or her consent, if it is not obvious that it is a montage or if it is not specified that it is a montage (WM)
  - To misuse / usurp someone else's identity (WM)



T2.1 – Definition of hate speech

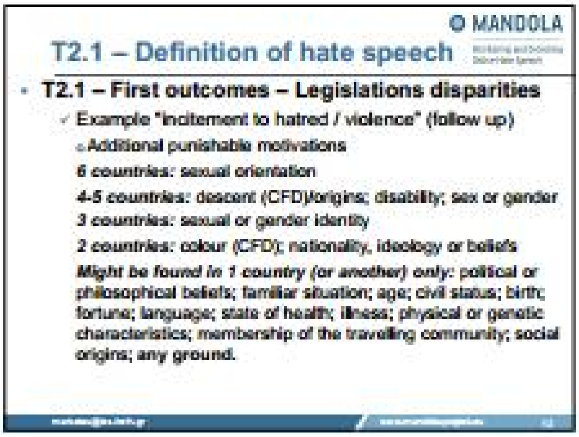
- **T2.1 – First outcomes**
  - **Important disparities between legislations:**
    - ✓ Most of transpositions of international and European instruments have not been done the same way, in addition to the fact that most countries provide for additional prohibitions
    - ✓ Example (one of the most common offence):

  Incitement to hatred or violence, directed against a group of persons / a person determined on the basis of their race, national or ethnic origin, and religion", (*eventually*) if the incitement is either carried out in a manner likely to disturb public order or is threatening, abusive or insulting (CFD 2008/913/JHA)



T2.1 – Definition of hate speech

- **T2.1 – First outcomes – Legislations disparities**
  - ✓ Example "incitement to hatred / violence" *(follow up)*
    - ○ 10 countries criminalise the incitement to hatred, but only 8 of them criminalise the incitement to violence;
    - ○ 8 countries additionally criminalise the incitement to discrimination (not mentioned in the CFD, but in the International Convention and in the additional protocol to the Convention on cybercrime)
    - ○ Only 3 countries impose an additional condition, and therefore only prohibit the public incitement to hatred if it is either carried out in a manner likely to disturb public order (2 countries), or public peace (1 country), or if it is threatening, abusive or insulting (1 country, alternatively to the disturbing of public order).



T2.1 – Definition of hate speech

- **T2.1 – First outcomes – Legislations disparities**
  - ✓ Example "incitement to hatred / violence" (follow up)
    - ○ Additional punishable motivations

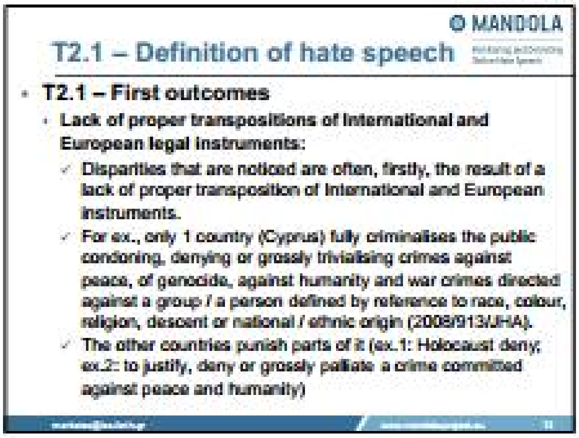
    6 countries: sexual orientation

    4-5 countries: descent (CFD)/origins; disability; sex or gender

    3 countries: sexual or gender identity

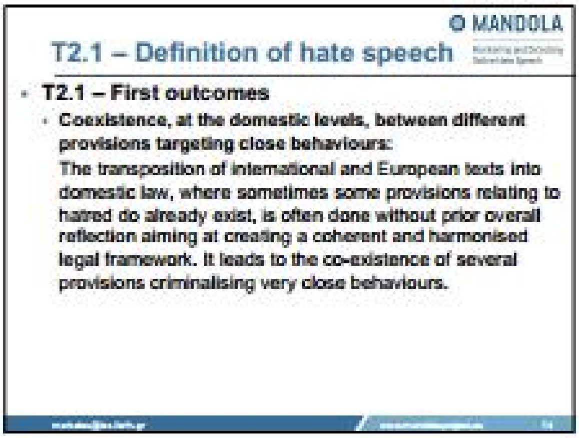
    2 countries: colour (CFD); nationality, ideology or beliefs

    Might be found in 1 country (or another) only: political or philosophical beliefs; familiar situation; age; civil status; birth; fortune; language; state of health; illness; physical or genetic characteristics; membership of the travelling community; social origins; any ground.



T2.1 – Definition of hate speech

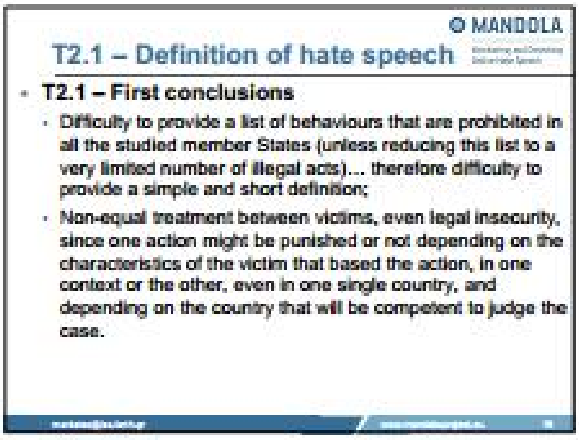
- **T2.1 – First outcomes**
  - **Lack of proper transpositions of international and European legal instruments:**
    - ✓ Disparities that are noticed are often, firstly, the result of a lack of proper transposition of international and European instruments.
    - ✓ For ex., only 1 country (Cyprus) fully criminalises the public condoning, denying or grossly trivialising crimes against peace, of genocide, against humanity and war crimes directed against a group / a person defined by reference to race, colour, religion, descent or national / ethnic origin (2008/913/JHA).
    - ✓ The other countries punish parts of it (ex.1: Holocaust deny; ex.2: to justify, deny or grossly palliate a crime committed against peace and humanity)



T2.1 – Definition of hate speech

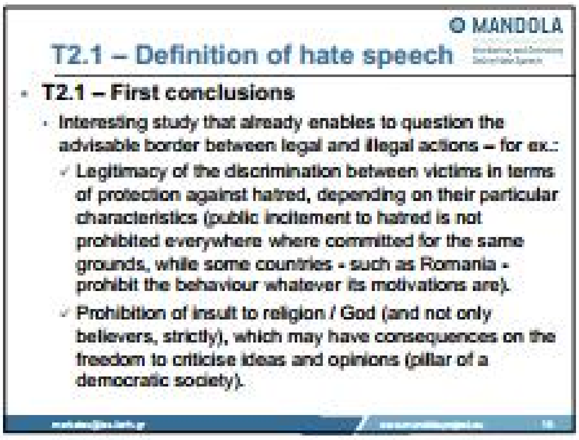
- **T2.1 – First outcomes**
  - **Coexistence, at the domestic levels, between different provisions targeting close behaviours:**

    The transposition of international and European texts into domestic law, where sometimes some provisions relating to hatred do already exist, is often done without prior overall reflection aiming at creating a coherent and harmonised legal framework. It leads to the co-existence of several provisions criminalising very close behaviours.



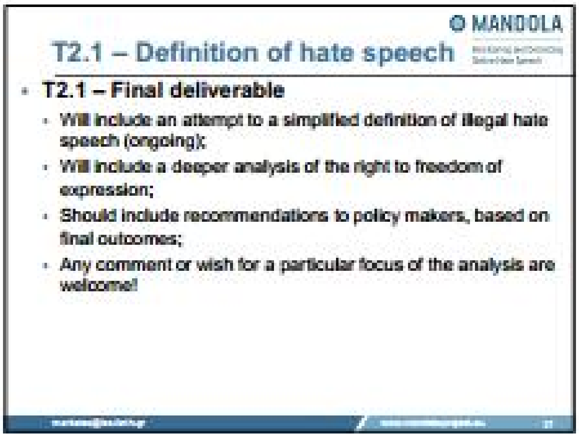
T2.1 – Definition of hate speech

- **T2.1 – First conclusions**
  - Difficulty to provide a list of behaviours that are prohibited in all the studied member States (unless reducing this list to a very limited number of illegal acts)… therefore difficulty to provide a simple and short definition;
  - Non-equal treatment between victims, even legal insecurity, since one action might be punished or not depending on the characteristics of the victim that based the action, in one context or the other, even in one single country, and depending on the country that will be competent to judge the case.



T2.1 – Definition of hate speech

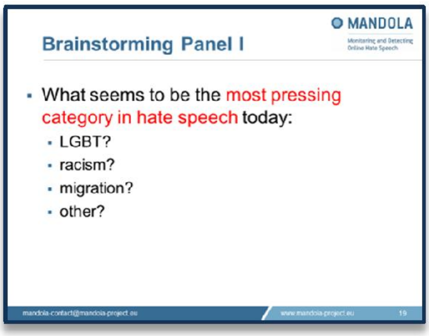
- **T2.1 – First conclusions**
  - Interesting study that already enables to question the advisable border between legal and illegal actions – for ex.:
    - ✓ Legitimacy of the discrimination between victims in terms of protection against hatred, depending on their particular characteristics (public incitement to hatred is not prohibited everywhere where committed for the same grounds, while some countries - such as Romania - prohibit the behaviour whatever its motivations are).
    - ✓ Prohibition of insult to religion / God (and not only believers, strictly), which may have consequences on the freedom to criticise ideas and opinions (pillar of a democratic society).



T2.1 – Definition of hate speech

- **T2.1 – Final deliverable**
  - Will include an attempt to a simplified definition of illegal hate speech (ongoing);
  - Will include a deeper analysis of the right to freedom of expression;
  - Should include recommendations to policy makers, based on final outcomes;
  - Any comment or wish for a particular focus of the analysis are welcome!

11 Appendix F: Brainstorming Panel I / Question 1

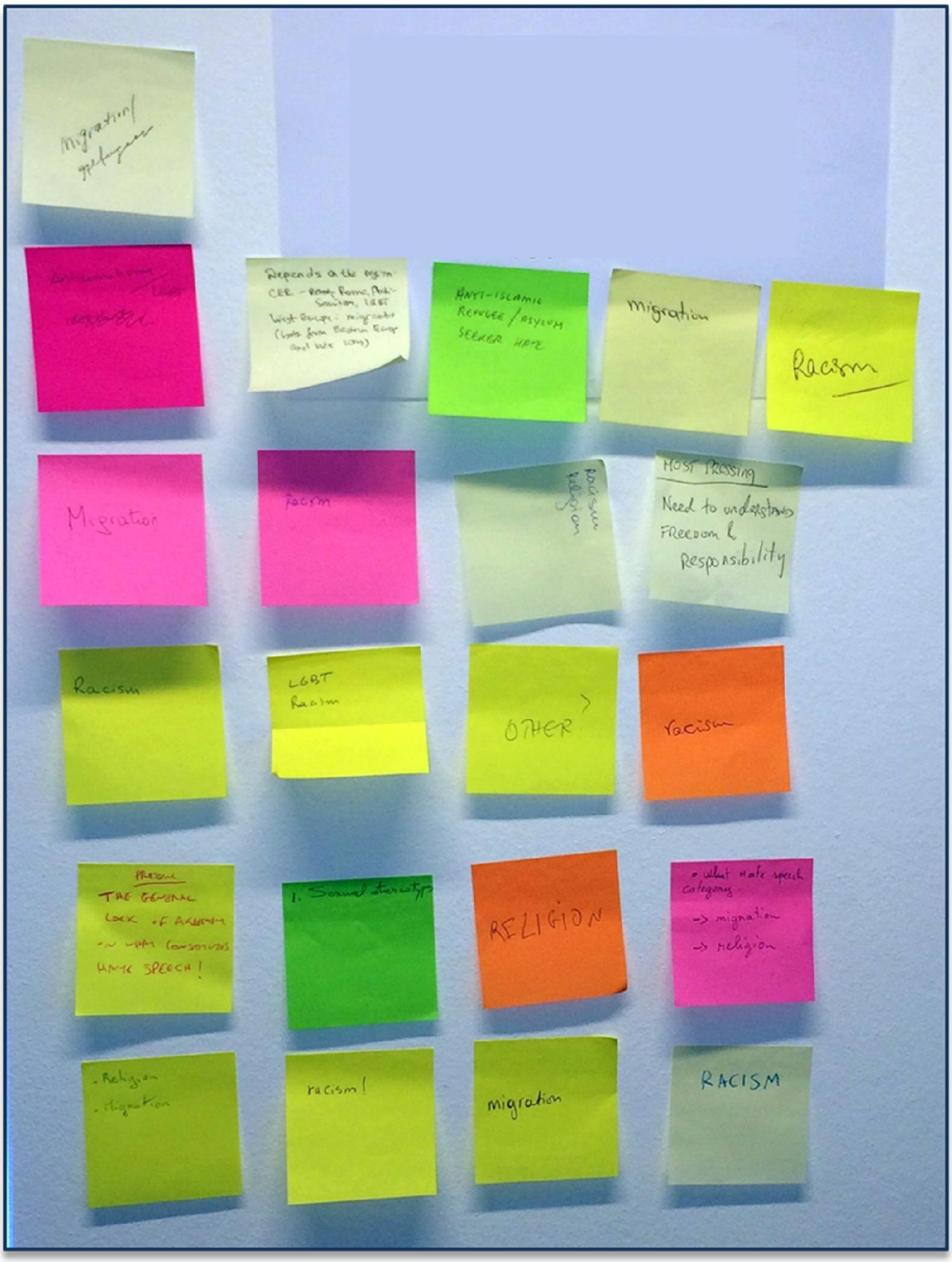


MANDOLA
Monitoring and Detecting Online Hate Speech

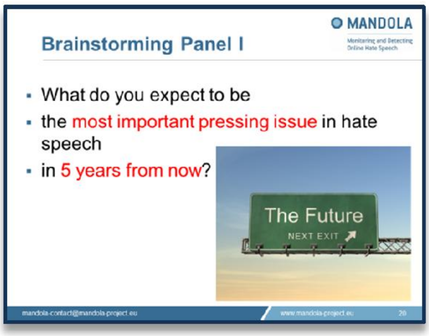
Brainstorming Panel I

- What seems to be the most pressing category in hate speech today:
  - LGBT?
  - racism?
  - migration?
  - other?

mandola-contact@mandola-project.eu www.mandola-project.eu 19



12 Appendix G: Brainstorming Panel I / Question 2

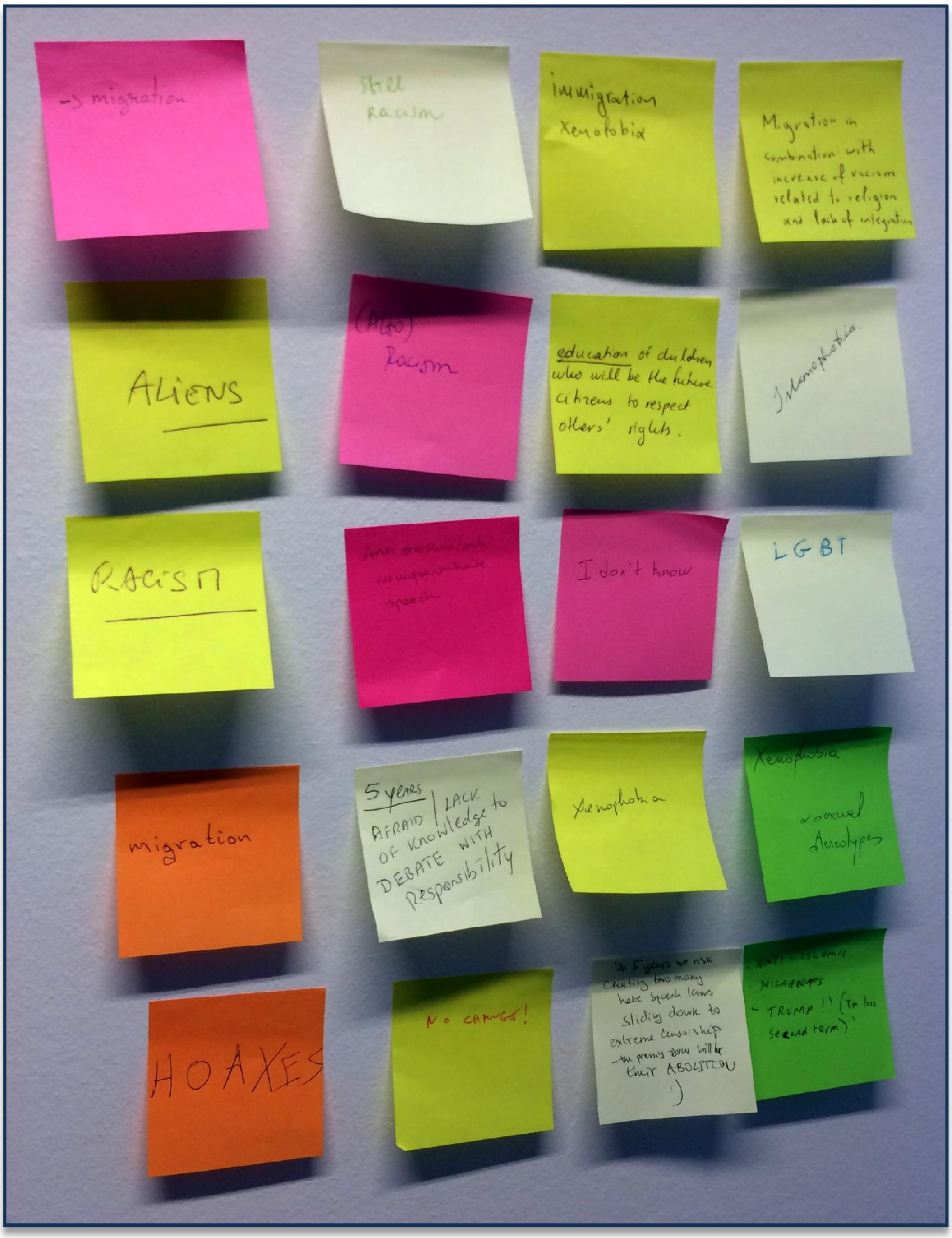


MANDOLA
Monitoring and Detecting Online Hate Speech

Brainstorming Panel I

- What do you expect to be
- the most important pressing issue in hate speech
- in 5 years from now?

mandola-contact@mandola-project.eu www.mandola-project.eu 20



-> migration

Still Racism

immigration
Xenofobia

Migration in combination with increase of racism related to religion and lack of integration

ALIENS

(ALSO) Racism

education of children who will be the future citizens to respect others' rights.

Islamophobia.

RACISM

Anti muslim/anti immigrant hate speech

I don't know

LGBT

migration

5 YEARS
AFRAID | LACK of Knowledge to DEBATE WITH Responsibility

Xenophobia

Xenophobia
sexual stereotypes

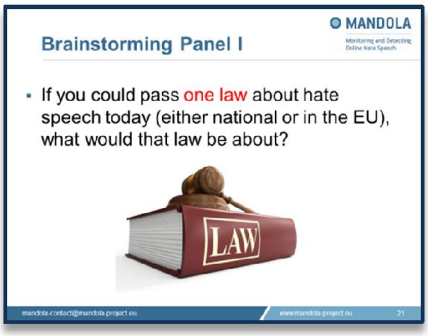
HOAXES

NO CHANGE!

In 5 years we risk creating too many hate speech laws sliding down to extreme censorship – the pressing issue will be their ABOLITION :)

- ANTI – ISLAMIC
- MIGRANTS
- TRUMP !! (In his second term)

13 Appendix H: Brainstorming Panel I / Question 3

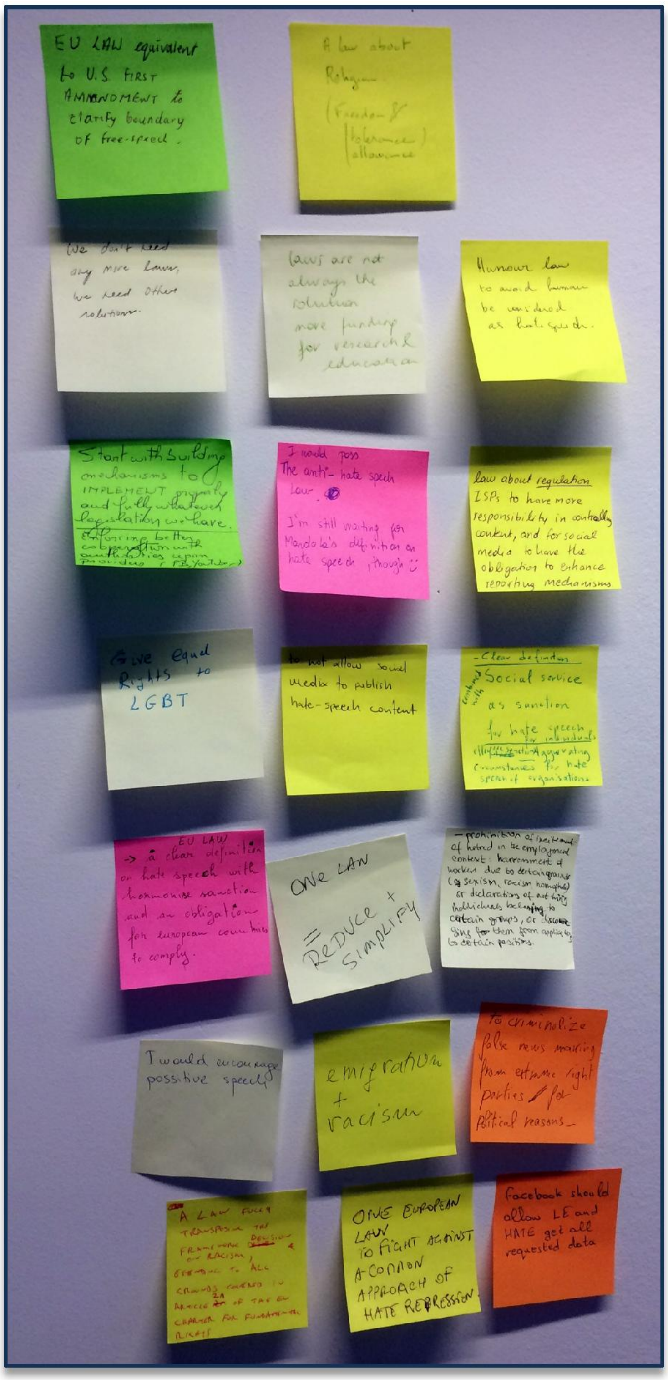


MANDOLA
Monitoring and Detecting Online Hate Speech

Brainstorming Panel I

- If you could pass one law about hate speech today (either national or in the EU), what would that law be about?

mandola-contact@mandola-project.eu
www.mandola-project.eu
21



EU LAW equivalent to U.S. FIRST AMMENDMENT to clarify boundary of free-speech.

A law about Religion (Freedom & tolerance allowance)

We don't need any more laws, we need other solutions.

laws are not always the solution more funding for research & education

Humour law to avoid humour be considered as hate speech.

Start with building mechanisms to IMPLEMENT properly and fully whatever legislation we have.

Enforcing better cooperation with authorities upon providers (FB, YouTube)

I would pass The anti-hate speech law.

I'm still waiting for Mandola's definition on hate speech, though :)

law about regulation ISPs to have more responsibility in controlling content, and for social media to have the obligation to enhance reporting mechanisms

Give equal Rights to LGBT

to not allow social media to publish hate-speech content

- Clear definition

combined with Social service as sanction for hate speech for individuals

[illegible] aggravating circumstances for hate speech of organisations

EU LAW → a clear definition on hate speech with harmonise sanction and an obligation for european countries to comply.

ONE LAW = REDUCE + SIMPLIFY

- prohibition of incitement of hatred in the employment context: harrassment of workers due to certain grounds (eg sexism, racism homophobia) or declarations of not hiring individuals belonging to certain groups, or discouraging them from applying to certain positions.

I would encourage possitive speech

emigration + racism

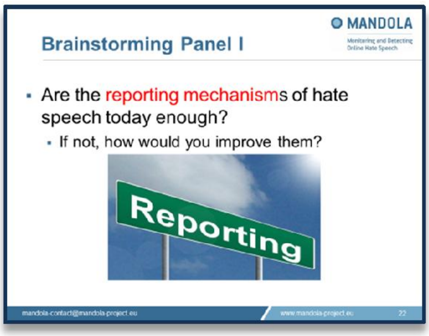
To criminalize false news making from extreme right parties / for political reasons.

A LAW FULLY TRANSPOSING THE FRAMEWORK DECISION ON RACISM, EXTENDING TO ALL GROUNDS COVERED IN ARTICLE 21 OF THE EU CHARTER FOR FUNDAMENTAL RIGHTS

ONE EUROPEAN LAW TO FIGHT AGAINST A COMMON APPROACH OF HATE REPRESSION.

Facebook should allow LE and HATE get all requested data

14 Appendix I: Brainstorming Panel I / Question 4

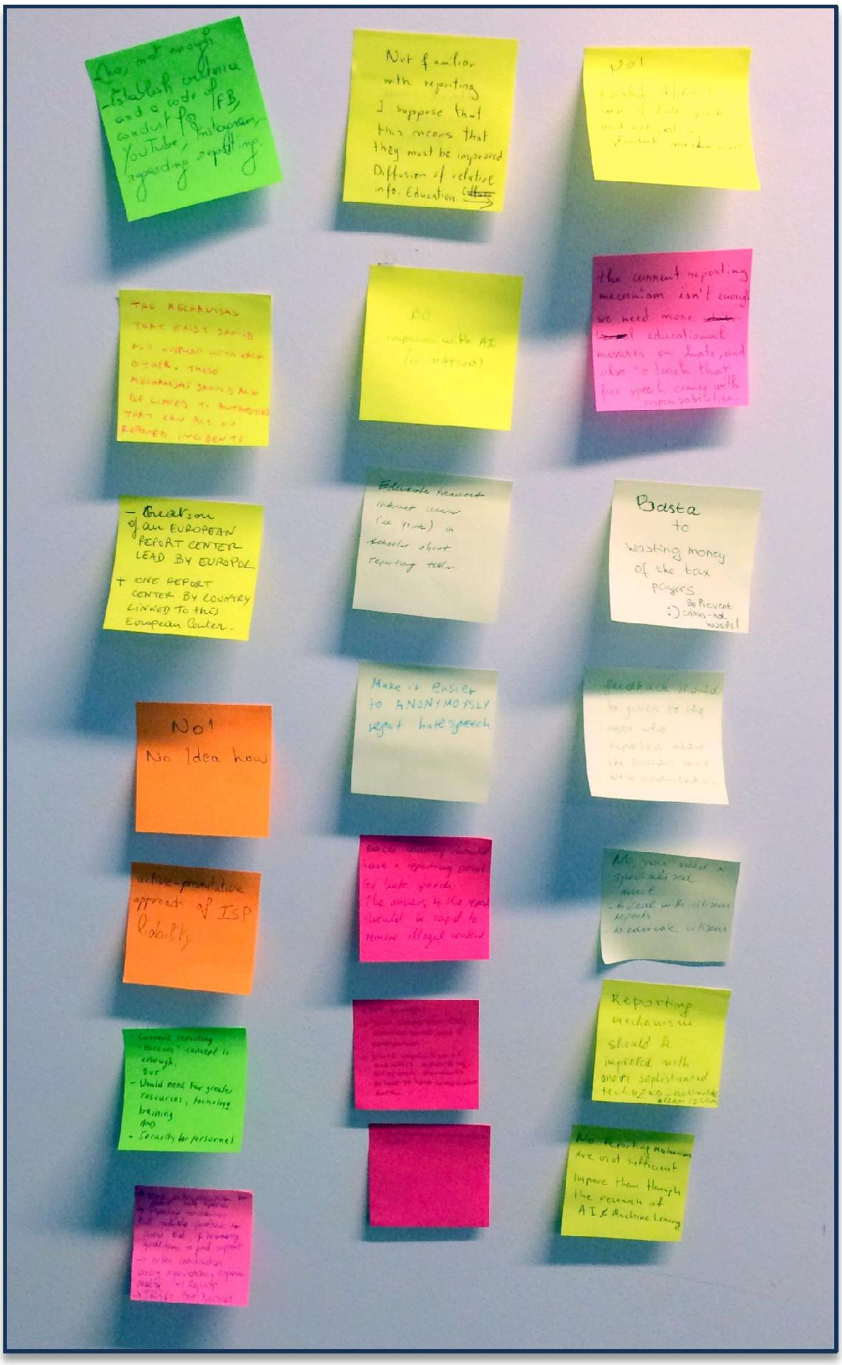


MANDOLA
Monitoring and Detecting Online Hate Speech

Brainstorming Panel I

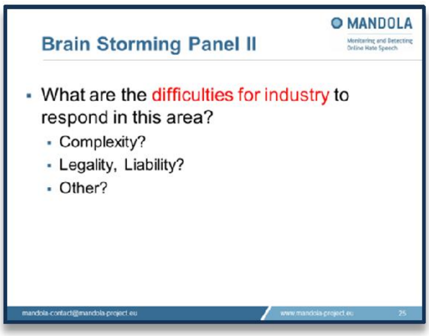
- Are the reporting mechanisms of hate speech today enough?
  - If not, how would you improve them?

mandola-contact@mandola-project.eu www.mandola-project.eu 22



No, not enough
- Establish criteria and a code of conduct for FB, YouTube, Instagram, regarding reporting.
Not familiar with reporting
I suppose that this means that they must be improved
Diffusion of relative info. Education
No!
THE MECHANISMS THAT EXIST SHOULD NOT OVERLAP WITH EACH OTHER. THESE MECHANISMS SHOULD ALSO BE LINKED TO AUTHORITIES THAT CAN ACT ON REPORTED INCIDENTS
NO
improve with AI
the current reporting mecanism isn't enough we need more educational mesures on hate, and also to teach that free speech comes with responsabilities.
- Creation of an EUROPEAN REPORT CENTER LEAD BY EUROPOL
+ ONE REPORT CENTER BY COUNTRY LINKED TO this European Center.
Educate teachers internet users (ie youth) in schools about reporting tools
Basta to wasting money of the tax payers
:) Prevent crimes not words!
No!
No Idea how
Make it easier to ANONYMOYSLY report hate speech
Feedback should be given to the user who reported about the decision that were undertaken
active-proactive approach of ISP liability
Each country should have a reporting point for hate speech. The answers to the reports should be rapid to remove illegal content
No, you need a specialised unit
- to deal with citizens reports
- to educate citizens
- Current reporting "Notice" concept is enough.
But
- Would need for greater resources, technology training
and
- Security for personnel
Reporting mechanisms should be improved with more sophisticated technologies
No. Reporting mechanisms are not sufficient
Improve them through the research of AI & Machine Learning

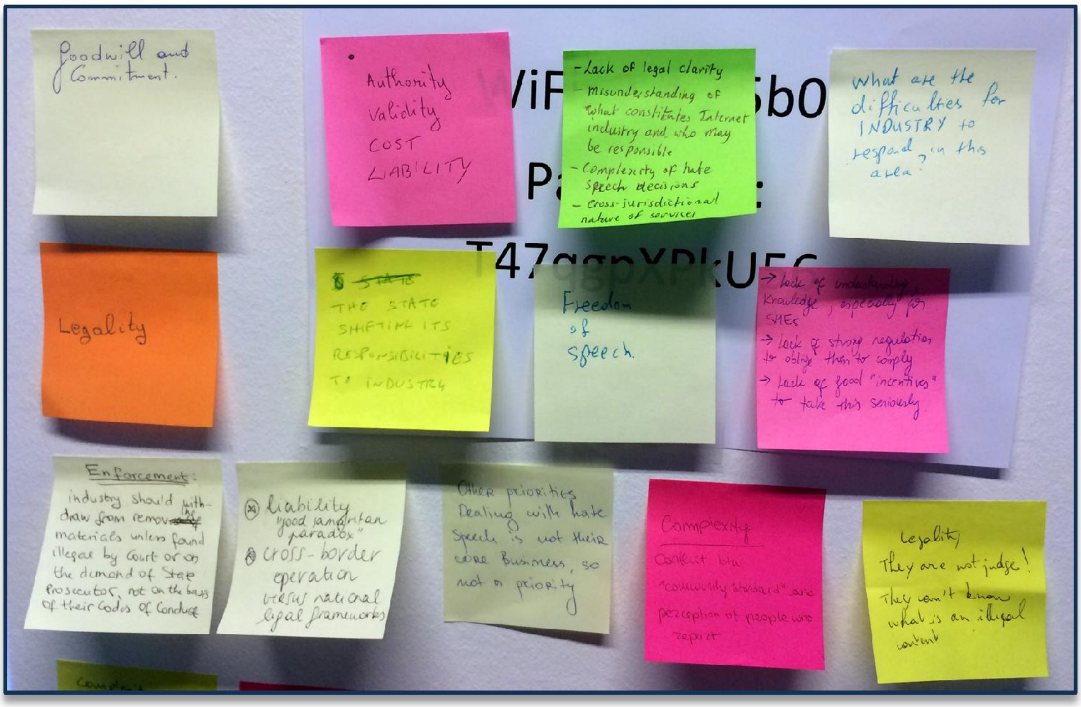
15 Appendix J: Brainstorming Panel II / Question 1



Brain Storming Panel II

- What are the difficulties for industry to respond in this area?
  - Complexity?
  - Legality, Liability?
  - Other?

mandola-contact@mandola-project.eu www.mandola-project.eu 25



Goodwill and Commitment.

Authority
Validity
COST
LIABILITY

- Lack of legal clarity
- misunderstanding of what constitutes Internet industry and who may be responsible
- Complexity of hate speech decisions
- Cross-jurisdictional nature of services

What are the difficulties for INDUSTRY to respond in this area?

Legality

~~STATE~~
THE STATE SHIFTING ITS RESPONSIBILITIES TO INDUSTRY

Freedom of speech.

→ lack of understanding, knowledge, especially for SMEs
→ lack of strong regulation to oblige them to comply
→ Lack of good "incentives" to take this seriously

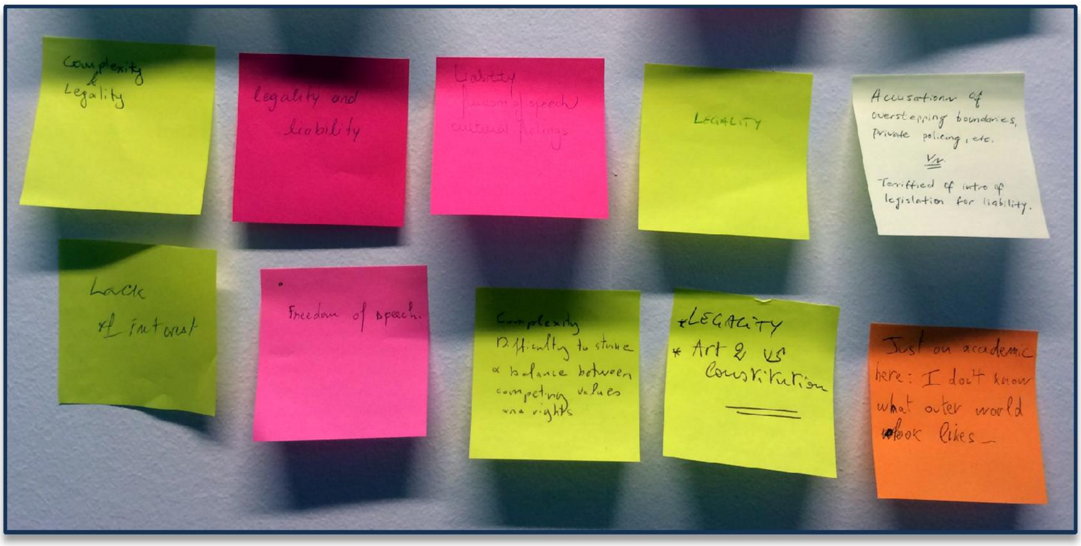
Enforcement:
industry should withdraw from removing materials unless found illegal by Court or on the demand of State Prosecutor, not on the basis of their Codes of Conduct

liability
"good samaritan paradox"
cross-border operation versus national legal frameworks

Other priorities
Dealing with hate speech is not their core business, so not a priority

Complexity
Conflict btw "community standard" and perception of people who report

Legality
They are not judge!
They can't know what is an illegal content



Complexity & Legality

legality and liability

Liability
Freedom of speech
cultural feelings

LEGALITY

Accusations of overstepping boundaries, private policing, etc.
Vs.
Terrified of intro of legislation for liability.

Lack of interest

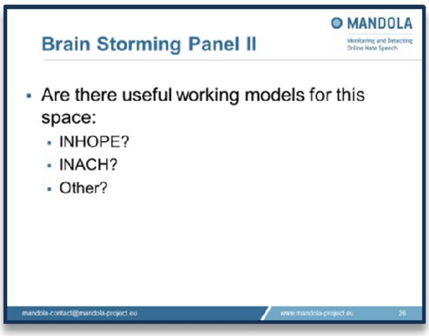
Freedom of speech.

Complexity
Difficulty to strike a balance between competing values and rights

*LEGALITY
* Art 2 vs Constitution

Just an academic here: I don't know what outer world look likes

16 Appendix K: Brainstorming Panel II / Question 2

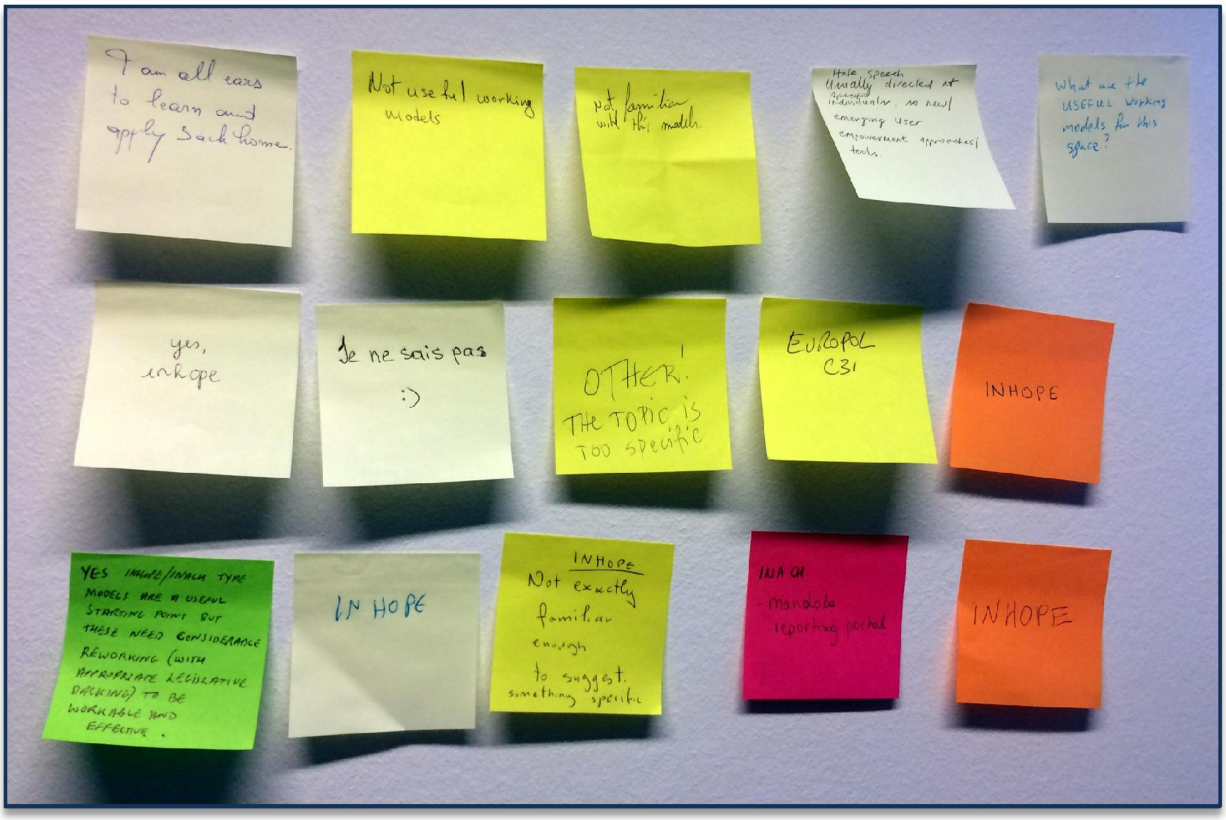


MANDOLA
Monitoring and Detecting Online Hate Speech

Brain Storming Panel II

- Are there useful working models for this space:
  - INHOPE?
  - INACH?
  - Other?

mandola-contact@mandola-project.eu www.mandola-project.eu 26



I am all ears to learn and apply back home.

Not useful working models

not familiar with the models.

Hate speech usually directed at specific individuals, no new/emerging user empowerment approaches/tools.

What are the USEFUL working models for this space?

yes, inhope

Je ne sais pas :)

OTHER! THE TOPIC IS TOO SPECIFIC

EUROPOL C3I

INHOPE

YES INHOPE/INACH TYPE MODELS ARE A USEFUL STARTING POINT BUT THESE NEED CONSIDERABLE REWORKING (WITH APPROPRIATE LEGISLATIVE BACKING) TO BE WORKABLE AND EFFECTIVE.

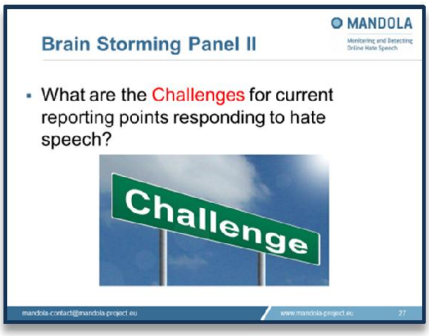
IN HOPE

INHOPE
Not exactly familiar enough to suggest something specific

INACH
- mandola reporting portal

INHOPE

17 Appendix L: Brainstorming Panel II / Question 3

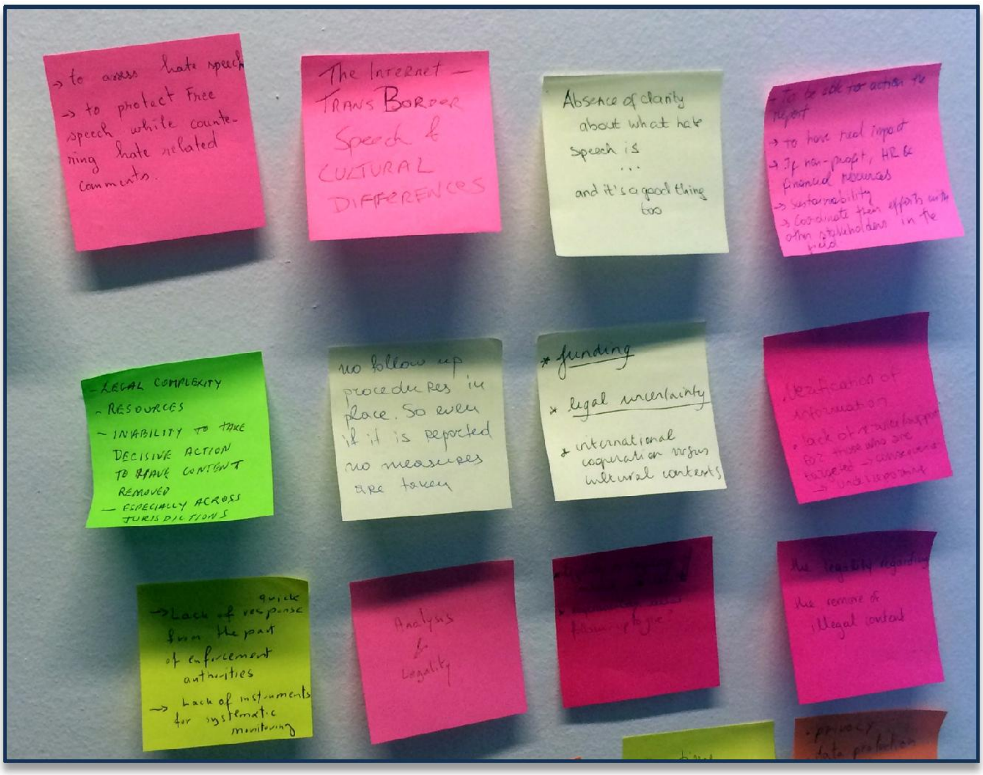


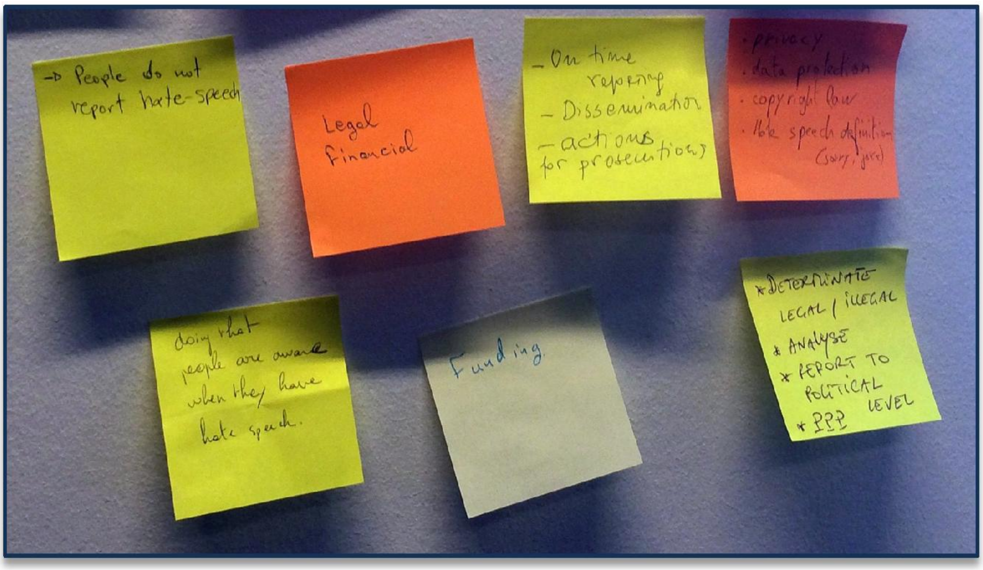
MANDOLA
Monitoring and Detecting Online Hate Speech

Brain Storming Panel II

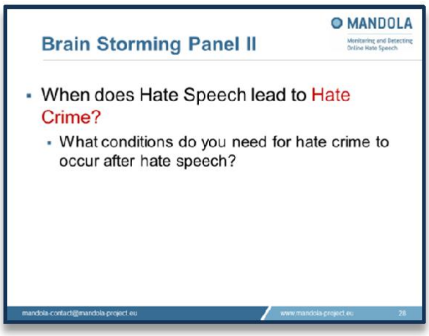
- What are the Challenges for current reporting points responding to hate speech?

mandola-contact@mandola-project.eu   www.mandola-project.eu   27


-> to asess hate speech
-> to protect Free speech while countering hate related comments.
The Internet – TRANS BORDER Speech & CULTURAL DIFFERENCES
Absence of clarity about what hate speech is ... and it's a good thing too
-> To be able to action the report
-> To have real impact
-> If non-profit, HR & financial resources
-> Sustainability
-> Coordinate their efforts with other stakeholders in the field
- LEGAL COMPLEXITY
- RESOURCES
- INABILITY TO TAKE DECISIVE ACTION TO HAVE CONTENT REMOVED
- ESPECIALLY ACROSS JURISDICTIONS
no follow up procedures in place. So even if it is reported no measures are taken
* funding
* legal uncertainty
* international cooperation versus cultural contexts
Verification of information
lack of ... support for those who are targeted -> consequences -> underreporting
-> Lack of quick response from the part of enforcement authorities
-> Lack of instruments for systematic monitoring
Analysis & Legality
the legality regarding the remove of illegal content
privacy
data protection


-> People do not report hate-speech
Legal
financial
_ On time reporting
_ Dissemination
_ actions for prosecution
. privacy
. data protection
. copyright law
. Hate speech definition (scary, jure)
doing that people are aware when they have hate speech.
Funding
* DETERMINATE LEGAL / ILLEGAL
* ANALYSE
* REPORT TO POLITICAL LEVEL
* PPP

18 Appendix M: Brainstorming Panel II / Question 4

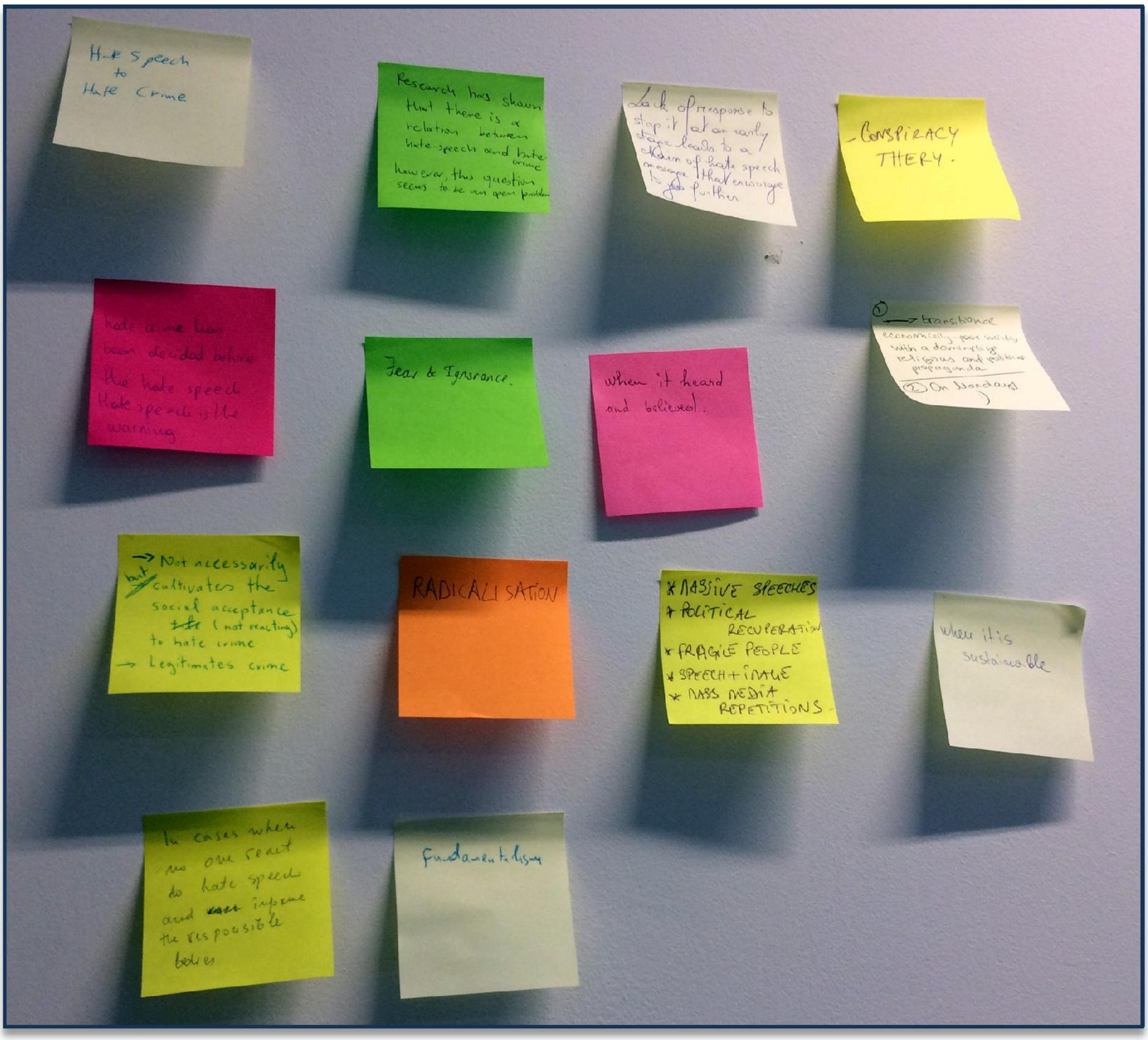


MANDOLA
Monitoring and Detecting Online Hate Speech

Brain Storming Panel II

- When does Hate Speech lead to Hate Crime?
  - What conditions do you need for hate crime to occur after hate speech?

mandola-contact@mandola-project.eu www.mandola-project.eu 28



Hate Speech
to
Hate Crime

Research has shown that there is a relation between hate-speech and hate crime however, this question seems to be an open problem

Lack of response to stop it at an early stage leads to a chain of hate speech message that encourage to go further

- CONSPIRACY THERY.

hate crime has been decided before the hate speech Hate speech is the warning

Fear & Ignorance.

when it heard and believed.

① ⟶ transitional economically poor society with a dominating religious and political propaganda

② On Mondays

→ Not necessarily ~~but~~ but cultivates the social acceptance ~~the~~ (not reacting) to hate crime
→ Legitimates crime

RADICALISATION

* MASSIVE SPEECHES
* POLITICAL RECUPERATION
* FRAGILE PEOPLE
* SPEECH + IMAGE
* MASS MEDIA REPETITIONS.

when it is sustainable

In cases when no one react to hate speech and ~~[illegible]~~ inform the responsible bodies

Fundamentalism