Rights, Equality and Citizenship (REC)

Programme of the EU Commission

(2014-2020)

Monitoring and Detecting Online Hate Speech

D2.4b (final report)

Privacy Impact Assessment of the MANDOLA outcomes

Abstract: The current report provides an account of the results of the privacy impact assessment (PIA) of the MANDOLA outcomes, which has been performed based on the method identified in D2.4a (intermediate).

Contractual Date of Delivery	30 September 2017
Actual Date of Delivery	30 September 2017
Deliverable Security Class	Public
Editor	Estelle De Marco
Quality and Ethical Assurance	Tatiana Synodinou and Cormac Callanan

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
UMO	Principal Contractor	France

Document Revisions & Quality Assurance

Internal Reviewers:

Tatiana Synodinou (UCY) (Chair of the Ethics Committee) and Cormac Callanan (AIS).

Revisions

Version	Date	By	Overview
v.2.4b.0	02/08/2017	Inthemis (FR) Estelle De Marco as editor	Preparation of the report, including technical contributions from George Pallis (UCY), Demetris Paschalides (UCY) and Álvaro Ortigosa (UAM).
v.2.4b.1	08/08/2017	Inthemis (FR) Estelle De Marco as editor	Second reading of the report including comments from Ioannis Inglezakis (Aristotle University / FORTH).
v.2.4b.2	10/08/2017	Inthemis (FR) Estelle De Marco as editor	Modifications following remarks from Tatiana Synodinou (UCY) (Chair of the Ethics Committee) and Cormac Callanan (AIS) as quality reviewers.
v2.4b.3	26/09/2017	Inthemis (FR) Estelle De Marco as editor	Preparation of the final report including new Sections 3.6 and 5 and modifying several sections of the report following the consultation of the members of the MANDOLA Advisory Board.
v2.4b.4	30/09/2017	Inthemis (FR) Estelle De Marco as editor	Final version after review by Tatiana Synodinou (UCY) (Chair of the Ethics Committee) as quality reviewer.

Table of Contents

List of Tables

1 Executive summary

The current report implements the privacy impact assessment (PIA) of the four main MANDOLA outcomes, which include a web-based monitoring dashboard, a smartphone app, a web-based reporting portal, and information dedicated to Internet users, the industry and policy makers.

As a consequence, it provides recommendations of the use and of the further development of these outcomes, with the aim of ensuring an appropriate protection of the right to private life and to personal data protection, and more widely of the other rights and freedoms either exercised by individuals in their respective personal spheres, or restricted by extension because of a privacy limitation or a personal data use (or non-use). Recommendations are summarised in Section 4.

The current PIA has been performed based on the method proposed in the MANDOLA Deliverable D2.4a.

2 Introduction

2.1 Background to the MANDOLA project

MANDOLA (Monitoring ANd Detecting OnLine hAte speech) is a 24-months project co-funded by the Rights, Equality and Citizenship (REC) Programme of the European Commission, which aims at making a bold step towards improving the understanding of the prevalence and spread of online hate speech and towards empowering ordinary citizens to report hate speech.

2.1.1 MANDOLA objectives

The MANDOLA specific objectives are the following:

- To monitor the spread and penetration of online hate-related speech in the European Union (EU) and in the E.U. Member States using big-data approaches, while investigating the possibility to distinguish, among monitored contents, between potentially illegal hate-related speech and non-illegal hate-related speech;
- To provide policy makers with actionable information that can be used to promote policies for mitigating the spread of online hate speech;
- To provide ordinary citizens with useful tools that can help them deal with online hate speech irrespective of whether they are bystanders or victims;
- To transfer best practices among E.U. Member States;
- To set-up a reporting infrastructure that will enable the reporting of potentially illegal hate speech.

The MANDOLA project addresses the two major difficulties in dealing with online hate speech: the lack of reliable data and the poor awareness on how to deal with the issue. Indeed, it is difficult to find reliable data that can show detailed online hate speech trends (inter alia in terms of geolocation and in relation to the focus of hate speech). Moreover, available data generally do not distinguish between potentially illegal hate speech and not illegal hate speech. In addition, the different legal systems in various Member States make it difficult for ordinary people to perceive the boundaries between both these categories of content. In this context, citizens might have difficulties to know how to deal with potentially illegal hate speech and how to behave when facing harmful but not illegal hate content. The lack of reliable data also prevents to make reliable decisions and push policies to the appropriate level.

The two MANDOLA innovations are (1) the extensive use of IT and big data to study and report online hate, and (2) the research on the possibility to make a clear distinction between legal and potentially illegal content taking into account the variations between E.U. Member States legislations.

MANDOLA is serving: (1) policy makers - who will have up-to-date online hate speech-related information that can be used to create enlightened policy in the field; (2) ordinary citizens - who will have a better understanding of what online hate speech is and how it evolves, and who will be provided with information for recognising legal and potentially

illegal online hate-speech and for acting in this regard; and (3) witnesses of online hate speech incidents - who will have the possibility to report hate speech anonymously.

2.1.2 MANDOLA activities

In order to achieve its objectives, the project includes the following activities:

- An analysis of the legislation on illegal hate-speech at the European and international level and in ten E.U. Member States.
- An analysis of the applicable legal and ethical framework relating to the protection of privacy, personal data and other fundamental rights in order to implement adequate safeguards during research and in the products to be developed.
- The development of a monitoring dashboard, which aims to identify and visualise cases of online hate-related speech spread on social media (such as Twitter) and on the Web.
- The creation of a multi-lingual corpus of hate-related speech based on the collected data, to be used to define queries in order to identify Web pages that may contain hate-related speech and to filter the tweets during the pre-processing phase. The vocabulary is developed with the support of social scientists and enhanced by the Hatebase (http://www.hatebase.org/).
- The development of a reporting portal, in order to allow Internet users to report potentially illegal hate-related speech material they have noticed on the Internet.
- The development of a smart-phone application, in order to allow anonymous reporting of potentially hate-related speech materials noticed on the Web and in social media.
- The creation and dissemination of a Frequently Asked Questions document, to be disseminated via the project portal and the smart-phone app.
- The creation of a network of National Liaison Officers (NLOs) of the participating Member States. They are intended to act as contact persons for their country, to exchange best practices and information, and to support the project and its activities with legal and technical expertise when needed.
- The development of a landscape of current responses to hate speech across Europe and of a Best Practices Guide for responding to online hate speech for Internet industry in Europe.

2.2 Purpose and scope of the report

The purpose of the current report is to present the results of the privacy impact assessment (PIA) of the MANDOLA outcomes, based on the method that has been proposed in the MANDOLA deliverable D2.1a.

This privacy impact assessment (PIA) has been conducted in relation with the four main MANDOLA outcomes, which are namely a monitoring dashboard, a smartphone app, a reporting portal, and information dedicated to Internet users, to policy makers and to the Internet industry. Its objective is to ensure that any further use of these outcomes will take place in a context of appropriate respect for the right to private life and to personal data protection, and more widely for the other rights and freedoms either exercised by individuals in their respective personal spheres, or restricted by extension because of a

privacy limitation or a personal data use (or non-use). To this end, the PIA includes recommendations of safeguards to be implemented by the MANDOLA consortium and by end-users.

This PIA has been performed based on the methodology proposed in the MANDOLA Deliverable D2.1b.

2.3 Document structure

The document is structured as follows.

Section 1 provides an executive summary.

Section 2 provides an introduction to the MANDOLA project and the current report.

Section 3 gives an account of the privacy impact assessment (PIA) that has been performed in relation to the four main MANDOLA outcomes.

Section 4 presents a summary of the PIA recommendations.

Section 5 provides a conclusion.

Section 6 provides a list of main acronyms and abbreviations.

Section 7 provides the list of the members of the MANDOLA Advisory Board who contributed to the privacy impact assessment.

3 Privacy impact assessment of the MANDOLA outcomes

Within the framework of the current privacy impact assessment (PIA), the notion of a PIA is understood in a broad sense in order to ensure an integral ethical approach, as including the assessment of risks posed by a project to the right to private life and to personal data protection, and more widely to the other rights and freedoms either exercised by individuals in their respective personal spheres, or restricted by extension because of a privacy limitation or a personal data use (or non-use)[1].

The method used for performing this PIA has been presented in the MANDOLA deliverable D2.4a (intermediate). It is divided into seven steps, and it is designed (1) to initiate the assessment "*as early as possible in the design of the processing operations*"[2], a need that has been recalled recently by the Article 29 data protection working party, and (2) to enable modification of the outcomes of previous steps in the course of the assessment itself. Indeed, a PIA is an iterative process, which means that each step should be (and has been, during the MANDOLA project) revisited and potentially "reworked" both during the course of the development of the project and during the course of the assessment. Firstly, this ensures privacy by design and by default, which is a requirement of the GDPR and the Directive on the protection of personal data in the justice and police sectors[3], by starting the PIA during the development of the project, "*even if some of the processing operations are still unknown*"[4]. Secondly, this enables the possibility to recognise, in the results of the first steps of the assessment, some elements (such as assets or supporting assets) that are only discovered during the performance of subsequent steps, and that might have consequences on the final results.

Some steps or sub-steps of this method do not receive precise answers, within the framework of the current assessment of the outcomes of the MANDOLA project, since the MANDOLA consortium does not control all the contextual elements relating to the organisations that will use these outcomes, nor the purposes for which these organisations might use and even further develop these project products. However, the MANDOLA consortium was determined to follow this method as accurately as possible, in order to ensure that the results of the assessment reflect the final project products potential impacts in the best possible way.

In any case, the MANDOLA outcomes listed above should not be reused prior to the implementation of a new PIA, which should be a comprehensive full-scale one in case of practical implementation, in order to consider the privacy risks posed by their new specific use context, including their purpose of use. In the situation where these outcomes would be

[1] An example provided by the Article 29 Data Protection Working Party is the financial loss that could result from inaccurate billing or price discrimination, which may be caused by a personal data processing (Article 29 Data Protection Working Party, Opinion 04/2013, *op. cit*., p. 7). Another example could be a (even temporary) deprivacy of liberty due to an investigation targeting someone other than the perpetrator of a penal offence, opened on the basis of the processing of non-reliable personal data.

[2] Article 29 Data Protection Working Party, *Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679* (WP248), *op. cit.* p.13.

[3] Article 25 of the GDPR and article 20 of the Directive for the police and criminal justice sector.

[4] Article 29 Data Protection Working Party, *Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679* (WP248), *op. cit.* p.13.

further developed, another comprehensive full-scale PIA should be planned during research in order to validate the envisioned new version.

3.1 Determination of the necessity of a PIA and its scale

In order to determine if a PIA is necessary, the following questions must be answered:

- **Is a PIA legally mandatory?**

  In the state of the current legislation, a PIA is not legally mandatory.

- **Does the project present any privacy risks?**

  The MANDOLA outcomes are likely to generate privacy risks, especially since they enable online content monitoring and since they might influence policies in the field of responding to online hate speech. This is particularly relevant in the light of the analysis of the legal and ethical context presented in the MANDOLA deliverable D2.2[5] and of the risks posed by the technologies that are used, which has been highlighted in Section 3.2.1 of the MANDOLA deliverable D2.4a (intermediate)[6].

  Issues at stake include the following:

  - Web scanning, including social networks scanning, may be seen as a disproportionate interference into the right to privacy. Indeed, individuals who publish information on the Internet in several different contexts expect the respect of each of these contexts. In other words, they do not expect and consent neither to the collection of their pieces of information out of their original context, nor to the combination of all their published information, all contexts taken together. Moreover, such a collection and comparison of information may lead to the creation of information on a given individual, without his or her knowledge and consent, which may lead to a high transparency of this individual vis-à-vis the data's controller or any person authorised to access these data.

    On the same line, victims of hate speech might want to reduce the visibility of the speeches referring to them. However if appropriate safeguards are implemented they might on the opposite value the objective of combatting these speeches, provided that measures taken in this regard actually serve their interests[7].

  - Online scanning of Tweets and webpages, in order to assess the potential illegality of their content and show the results in a dashboard, may imply by default the recording of a high number of data, while the E.U. legal instruments protecting personal data require, on the opposite, a minimisation of the collected data.

  - The storage of Tweets URLs, and even the capacity of search engines to trace back a public Tweet on the basis of some keywords of its content, enables in practice the direct or indirect identification of the author of this specific Tweet. Access to this

[5] MANDOLA Deliverable D2.2 - *Identification and analysis of the legal and ethical framework*, version 2.2.4 of 12 July 2017, MANDOLA project (Monitoring ANd Detecting OnLine hAte speech) - GA n° JUST/2014/RRAC/AG/HATE/6652, http://mandola-project.eu/.

[6] MANDOLA Deliverable D2.4a (Intermediate) - *Privacy Impact Assessment of the MANDOLA outcomes*, version 2.4a.2 of 11 July 2017, MANDOLA project (Monitoring ANd Detecting OnLine hAte speech) - GA n° JUST/2014/RRAC/AG/HATE/6652, http://mandola-project.eu/.

[7] This paragraph has been added following consultation of the Mandola Advisory Board members.

information might be disproportionate and might present high risks for fundamental rights in certain contexts, particularly where the Tweet would have been assessed as being or containing potentially illegal content by a technical feature or a social analyst.

- In addition to the preceding note, the direct or indirect identification of an individual could be combined, in practice, with other data processing operations pursuing the same or a different purpose. Such interconnections might generate important privacy risks and would require - provided that they are not prohibited by law - specific safeguards and close monitoring.
- The production of statistics related to the number of illegal contents originated from one given city or country might lead to the stigmatisation of the citizens of this city or country, and therefore to discrimination, with negative impacts on several other fundamental rights[8].
- The sending of reports relating to potentially illegal content to specialised hotlines or to LEAs might raise concern related to the confidentiality of the identity of the reporting person, of the alleged perpetrator of a potential penal offence and of the victim of this offence. It also might lead to the commencement of legal proceedings, and/or to Internet private censorship, and might impact on several protected freedoms such as the right to liberty and the right to a fair trial. For these reasons it must be protected with specific technical, organisational and legal safeguards.
- Recommendations relating to the legal definition of hate speech and relating to the instruments for combating this hate-speech phenomenon might lead to policies decisions which will impact individuals. For this reason they must be very carefully developed.

Therefore, the necessity to conduct a PIA is validated.

This PIA must be a full-scale PIA, given the potentially significant impacts described above.

3.2 Determination of the assessment team and of its objectivity

The privacy impact assessment team is the legal and ethical research consortium as described in the MANDOLA proposal in relation with Work stream 2, led by Inthemis. The members of this team, which are reasonably independent from potential end-users, have done their best efforts to recognise their potential subjectivity in this assessment report, in addition to justifying each of the decisions they take. They have been chosen for their specific expertise and personal independence.

The method adopted in order to implement the current PIA has already been chosen and described in the MANDOLA Deliverable D2.4a.

The decision to implement a PIA has two origins. The members of the assessment team have recognised the necessity to conduct a PIA during the previous step of the current PIA, but the decision to conduct this assessment has also been made within the framework of the description of the MANDOLA proposal, as a commitment of the MANDOLA consortium. The result has been the definition of a task 2.4, within WP2, entitled "Privacy impact assessment of the operational system (intermediate and final)". Therefore, the resources allocated to

[8] Including impact on tourism and trading activities.

the PIA and the PIA schedule have been determined at the origin of the project, and appeared to be appropriate.

During the assessment and more widely during the entire cycle of the MANDOLA research, legal experts and technical experts have exchanged information through teleconferences, physical meetings and emails involving either a subset of interested partners or involving all the MANDOLA partners, depending on the level of interest in the discussion by the whole consortium. These exchanges have enabled a successful communication between partners and the performance of an efficient way of working, both on and from the technical and the legal/ethical sides.

Regarding the approval of the final results of this PIA, the process has been as follows:

- The first version of the PIA results has been included in a preparatory MANDOLA deliverable D2.4b, which has been validated by all the MANDOLA partners and has been subjected to a quality and ethical review before being sent to the Advisory Board (AB) members in order to collect their opinions on first recommendations (see step 3.6 of the current PIA);
- The final version of Deliverable D2.4b, including a summary of the AB review and possibly new recommendations resulting from the AB consultation, will be firstly reviewed internally and validated by the MANDOLA partners. Afterward, the deliverable will be the subject of an additional quality and ethical review, before being submitted to the European Commission.

The manner in which the PIA recommendations have been implemented, where such implementation were feasible during the MANDOLA research, has been decided by the MANDOLA partners. Privacy safeguards that could not be implemented during the MANDOLA research (mostly due to the research limits or because only end-users can implement them) are the subject of recommendations of product use or of further development after the end of the MANDOLA project.

3.3 Description of the scope and framework of the study

Following the method described in D2.4a, the current PIA step aims at determining all the "*basic parameters within which risks must be managed*"[9], in other words all the elements that are necessary to the risk management step[10], namely a description of the framework of the study (which enables to ensure the feasibility of the study and to orientate works and deliverables based on real objectives), the identification of the assets (which refer to the assets, including immaterial, that need to be protected, and to the supports of these assets - human, hardware, software...), and a preparation of metrics (which are the parameters and scales which will be used to manage risks).

[9] This sentence is from the ENISA: ENISA website, Risk Management, Definition of scope and framework, available at http://www.enisa.europa.eu/activities/risk-management/current-risk/risk-management-inventory/rm-process/crm-strategy/scope-framework (home/our activities/risk management/current risk/ rm inventory/rm process/crm strategy/Scope & Framework), last accessed on 15 June 2017.

[10] The risk management step includes the analyses presented in Section 3.4.

3.3.1 Description of the framework of the study

This section includes a description of the frame of the study, a description of the context of the study, a description of the envisioned products or processing operations, a description of the scope and boundaries of the study, an identification of the parameters to be considered and an identification of the threat sources.

3.3.1.1 Description of the frame of the study

The objective of the study is to perform a PIA of the MANDOLA main outcomes, which are namely the following:

- to create a web-based monitoring dashboard aiming to monitor the spread and penetration of on-line hate-related speech in the European Union (E.U.) and in the E.U. Member States using big-data research technologies;
- to create a smartphone app that will enable the reporting of potentially illegal online hate speech;
- to create a web-based reporting portal that will enable the reporting of potentially illegal hate speech;
- to provide information to:
    - policy makers, in order to assist them in the promotion of policies for mitigating the spread of hate speech online;
    - ordinary citizens, in order to help them respond to hate speech online irrespective of whether they are observers or victims;
    - the Internet industry, in order to assist them in the identification of best practices and challenges in responding to hate speech online.

The objective of the PIA is to accompany the delivery of these MANDOLA outcomes with recommendations for implementation, use and (if necessary) further development before implementation and use, in order to make the use of these outcomes compliant with legal and ethical rules protecting privacy, personal data and other fundamental rights.

However, it will be identified in further steps and sub-steps of the current PIA that it is not possible, during the MANDOLA project, to take into account all the elements of context that will surround the use of certain MANDOLA outcomes in the future. This is particularly true when the user organisations which will adopt these outcomes are unknown today. For this reason, the current PIA and the respect of its conclusions must not be considered as a *carte blanche* to reuse all the MANDOLA outcomes without further precautions: subsequent PIAs, taking into account the future developments of the technical systems and the specificities and environment of the organisations that will use these systems and other outcomes might have to be performed by independent experts or by end-users under the control of such experts. Control of the actual complete and comprehensive implementation of the PIAs' results must also be ensured and audited.

The expected deliverable is Deliverable D2.4b, entitled *Privacy Impact Assessment of the MANDOLA outcomes*. First results of the PIA have been included in an intermediate Deliverable D2.4b, which has been transmitted to the Advisory Board members, in order to collect their views on first PIA recommendations with the objective of taking these views into account to the utmost extent within the framework of the PIA's final results.

Regarding the procedure that has been followed, this PIA has been performed by the legal and ethical consortium, and has been validated as explained in Section 3.2 of the current report.

3.3.1.2 Description of the context of the study

The context of this study is the MANDOLA project, which is a two-years project ending in September 2017, part funded by the Rights, Equality and Citizenship (REC) Programme of the European Commission.

The MANDOLA project is about improving our understanding of the prevalence and spread of online hate speech, empowering citizens with reporting tools and awareness information, empowering policy makers with actionable information that can be used to promote policies for mitigating the spread of on-line hate speech, and empowering the Internet industry with information on best practices and challenges in responding to hate speech online.

The MANDOLA outcomes are described in more details in Section 3.3.1.3.

All these products will not be ready for use at the end of the MANDOLA project. Indeed,

- The reporting portal as well as the information dedicated to Internet users, to policy makers and to the Internet industry will be available at the end of the project.
- The monitoring dashboard and the smartphone app will be delivered at TRL (Technology Readiness Level) 7: System prototype demonstration in an operational environment[11]. This means that further developments (including testing) are required before any commercialisation.

Regarding end-users of the MANDOLA outcomes, they are or might be the following:

- The MANDOLA partners and other private entities including the ISP industry.
- The general public as regards the reporting portal, the information dedicated to Internet users, the smartphone app and the monitoring dashboard.
- Policy makers as regards information dedicated to them and the monitoring dashboard.
- LEAs and assistance services against online hate speech as regards the smartphone app and the monitoring dashboard.

Therefore, the current PIA is not only about the implementation and/or use of the MANDOLA outcomes by the MANDOLA partners' organisations, but about the implementation and/or use of these outcomes, at the end of the MANDOLA project or after further development, by several persons and stakeholders, including law enforcement organisations. Moreover, at the end of the project, the MANDOLA partners might decide to remain involved in the MANDOLA website maintenance, and therefore in the diffusion of information dedicated to Internet users, to the Internet industry and to policy makers, in the making available of the reporting portal, and eventually in the diffusion of the monitoring dashboard results and of the smartphone app. The legal form under which this collaboration will take place, and the exact entities that will have the responsibility for the information storage, potential updates and diffusion, are currently unknown.

As a result, given the diversity of possible internal and external contexts of implementation and/or use of these outcomes, it is not possible to make a presentation, in the current

[11] More details about the TRLs can be found here https://en.wikipedia.org/wiki/Technology_readiness_level.

report, of all the possible internal and external contexts of use of these outcomes, for which a PIA is required. This is one of the arguments that supports the importance of performing other PIAs in the future, under subsequent end-users' (or information providers') responsibility.

Despite previous comments, it is important to attempt to define as precisely as possible the notion of risks in the final context of use of the MANDOLA outcomes.

In this regard, the MANDOLA research consortium retains a large definition of risks. The risks that are assessed within the framework of the current PIA - and that should at least be assessed within the context of subsequent PIAs - are the following:

- **As regards the monitoring dashboard and the smartphone app (and therefore in relation with computer systems) that have been developed by the MANDOLA consortium:**

  Any event, fact, or action,

  - caused by the system itself (by failure, defect or obsolescence) or by the computer system that supports or broadcasts it, or by another computer system connected to one of the latter systems, or by a software injected in one of the latter systems, or by a person having access to one of the latter systems, or by any other person even outside the organisation, or by a natural phenomenon;
  - directed against one of these systems, or against a person, an entity or a material belonging to the internal or external context of use of these systems, one of their components or softwares, or one or several data these systems contain;
  - having the effect (including as an indirect result) of interfering with private life or personal data protection or more widely with fundamental rights either exercised by individuals in their respective personal spheres, or restricted by extension because of a privacy limitation or a personal data use (or non-use)[12].

- **As regards the reporting portal and the information to be provided to Internet users, to the industry and to policy makers (and therefore in relation with any information) that has been created during the MANDOLA research and aiming at being broadcast after the project:**

  Any event, fact, or action,

  - caused by the information (by defect or obsolescence), or by the computer system that supports or broadcasts it, or by another computer system connected to the latter system, or by a software injected in this system, or by a person having access to this system or to the information, or by any other person even outside the organisation, or by a natural phenomenon;
  - directed against the provided information or its supporting computer system, or against a person, an entity or a material belonging to the internal or external context of diffusion or of use of the information or to the internal or external

[12] The analysis of the right to private life, of the right to personal data protection, and of the other fundamental rights at stake, is available in the MANDOLA deliverable D2.2. The explanation of the reason why these latter other fundamental rights are taken into account is available in the MANDOLA deliverable D2.4a, Section 3.1.1.

context of the information supporting system, one of the components or softwares of this latter system, or one or several data this system contains;

- having the effect (including as an indirect result) of interfering with private life or personal data protection or more widely with fundamental rights either exercised by individuals in their respective personal spheres, or restricted by extension because of a privacy limitation or a personal data use (or non-use)[13].

Regarding the organisation in the area of risk management of the end-users' structures, it is not possible to describe it in this report (beyond some details that will be provided in the section relating to the detailed description of the project), for the reasons exposed above.

3.3.1.3 Detailed description of the envisioned project or processing operations

The MANDOLA products subject to this PIA aim at contributing to the combat against hate speech online. Therefore, a detailed description of the envisioned project implies to evoke firstly the definition of "hate speech" that bases the MANDOLA works, before describing the products more precisely.

3.3.1.3.1 The notion of hate speech[14]

The determination of the contents to be targeted by the MANDOLA research has been the subject of the MANDOLA Deliverable D2.1 entitled *Definition of illegal hatred and implications*[15], of which the intermediate version (a) has been delivered in July 2016 and the final version (b) has been delivered in September 2017.

As explained in Section 6.1 of Deliverable D2.1b, it has been found necessary that the MANDOLA research focusses on illegal hate speech, defined as hate publications (covering all forms of written and oral expression) that are currently illegal. In order to give a precise content to illegal hate speech, a comparative study of ten E.U. Member States (namely Belgium, Bulgaria, Cyprus, France, Germany, Greece, Ireland, the Netherlands, Romania, and Spain) has been performed[16]. The scope of the investigation has been deliberately broad in order to perform an extensive analysis of legislations. It has covered all the penal provisions, along with civil or even administrative ones, that prohibit actions that have a link with hate, even if relatively thin, in other words in which the perpetrator demonstrates a particular intent to hurt or prejudice another person or group of persons, or to commit an action that is very likely to have such an effect. The personal characteristics of the victims that might motivate the perpetrator's action (such as religion, colour or gender) have not been taken into account as exclusion criteria, in order to investigate countries' choices in this area. In addition, some provisions have been knowingly ignored, since they have been considered to go too far beyond the core of the study, namely the processing of sensitive personal data and the provisions relating to audiovisual media services[17].

---

[13] See the preceding footnote.

[14] This Section has been added following consultation of the Mandola Advisory Board members.

[15] MANDOLA Deliverables D2.1a and D2.1b - *Definition of illegal hatred and implication*, September 2017, MANDOLA project (Monitoring ANd Detecting OnLine hAte speech) - GA n° JUST/2014/RRAC/AG/HATE/6652, http://mandola-project.eu/publications/.

[16] See the executive summary and Section 2.2.1 of the MANDOLA Deliverable D2.1b.

[17] See Section 2.3 of the MANDOLA Deliverable D2.1b.

Since this analysis has shown a wide heterogeneity and complexity of legislations, the choice has been made to classify illegal behaviours into four main categories: (1) behaviours that are illegal in all or almost all the E.U. Member States studied, (2) behaviours that are illegal or partly illegal in a majority of the E.U. Member States studied, (3) behaviours that are illegal in a minority of the E.U. Member States studied, and (4) additional behaviours that should be illegal according to European and International instruments[18].

In the three first of the above-mentioned categories, several illegal behaviours (covering the territory of the countries studied taken together) have been defined in their most common definition where found possible. Where it was not possible due to a too wide heterogeneity of legislations, illegal behaviours have been defined according to existing European and/or international instruments. Where there were no such instruments available, the retained definition has been the more interesting one in terms of "novelty" compared to other close illegal behaviours already studied[19]. All these findings have been reported in tables highlighting the specificities of each country compared to each "main" definition[20].

These three first categories of prohibited behaviours have served as a basis of the MANDOLA technical works, including the work of the social analysts who have contributed to the technical research[21].

3.3.1.3.2 Description of the MANDOLA outcomes

The project is mainly to deliver the four following outcomes already outlined in the previous sections, namely:

1. **A monitoring dashboard**

   The purpose of the monitoring dashboard is to monitor the spread and penetration of online hate-related speech in Europe and in Member States using big data approaches[22], in order to offer reliable information about online hate speech enabling users to focus on their geographic region ranging from their city to their country to the entire European Union[23].

   The monitoring dashboard is described in details in the MANDOLA deliverable D3.1 and it does not seem necessary to describe it again so deeply in the current report.

   Basically, the system is designed to dynamically analyse in real-time Tweets and web contents from Google streams, in order to extract from them only a smaller number of data elements that are namely the qualification of the content on a legal point of view (hate speech or non-hate speech), the type of hate speech, and the geolocation (timestamp of the tweets where the user has enabled the location service), reducing the decimals of the coordinates to 3 in order to minimise the possibility of any identification of

[18] See for example Section 2.2.2 of the MANDOLA Deliverable D2.1b.

[19] See Section 2.3 of the MANDOLA Deliverable D2.1b.

[20] See Section 7 of the MANDOLA Deliverable D2.1b.

[21] See Section 3.3.1.3.2 of the current report.

[22] MANDOLA Deliverable D3.1 - *MANDOLA Monitoring Dashboard*, MANDOLA project (Monitoring ANd Detecting OnLine hAte speech) - GA n° JUST/2014/RRAC/AG/HATE/6652, http://mandola-project.eu/, p. 7.

[23] MANDOLA Deliverable D3.1 - *MANDOLA Monitoring Dashboard*, *op. cit.* p. 8.

a particular individual. These elements are stored in a database and visualised in the dashboard.

In addition, a module called the "ground truth data set"[24] implies the storage of potentially illegal Tweets and texts coming from the Internet, along with their URL, in a "hate speech database", and their evaluation by human analysts. The function of the ground truth data set is to train the "text classifier", which is the component which classifies contents as legal or illegal[25]. During the MANDOLA research[26], the ground truth data set, "*built from* [a] [...] *sample set of tweets and Google pages that have been evaluated by social scientists*"[27], has been provided by UCY as an external service. This data set will not be included in the dashboard to be delivered.

However, after the MANDOLA research, the text classifier should ideally be further trained. This future training could be done by sending new texts to the ground truth data set proposed by UCY or to another data set provided in the research community in order to train classifiers for detecting hate speech[28]. This sending could be done by human analysts such as the moderators who analyse hate speech content as a part of their work.

Technically, this system and all these components can be hosted on the same server.

As regards the degree of finalisation of the monitoring dashboard, it will be delivered at TRL (Technology Readiness Level) 7: System prototype demonstration in an operational environment. This means that further developments (including testing) are required before commercialisation.

As regards information flows, the dashboard uses Twitter and Web sites as sources of possible hate-related online content[29], and other social networks could also be scanned. Neither content-related elements nor extracted words are kept. Twitter user mentions are also removed from the text before analysis. However, contents received from new human analysts in order to train the classifier could be kept by providers of the data set aiming to perform this training.

Information shown in the dashboard through the user interface is a hate map[30] and a hot-spot density[31] showing the percentage of suspected hate speech in each area. Results can be filtered by dates and by hate-speech categories. In addition, hate speech percentages

---

[24] MANDOLA Deliverable D3.1 - *MANDOLA Monitoring Dashboard*, *op. cit.* pp. 18 *et seq*.

[25] The assessment of the illegality of a given content is based on a) the classification algorithm and b) the ground truth data set. Specifically, UCY has used the Stochastic Gradient Descent classifier from the following library: http://scikit-learn.org/stable/modules/sgd.html#classification. The SGD Classifier has a number of parameters (see http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html#sklearn.linear_model.SGDClassifier). Using the Randomised search method (see http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html), the research team has searched for the best parameters for the classifiers (based on their data set).

[26] The analysis of the legal and ethical compliance of the MANDOLA research is available in the MANDOLA Deliverable D2.3 - *Legal and ethical compliance of research*, September 2017, MANDOLA project (Monitoring ANd Detecting OnLine hAte speech) - GA n° JUST/2014/RRAC/AG/HATE/6652, http://mandola-project.eu/.

[27] MANDOLA Deliverable D3.1 - *MANDOLA Monitoring Dashboard*, *op. cit.* p. 19.

[28] See https://data.world/crowdflower/hate-speech-identification.

[29] MANDOLA Deliverable D3.1 - *MANDOLA Monitoring Dashboard*, *op. cit.* p. 8.

[30] See the MANDOLA Deliverable D3.1 - *MANDOLA Monitoring Dashboard*, *op. cit.*, Section 7.1, p.24.

[31] See the MANDOLA Deliverable D3.1 - *MANDOLA Monitoring Dashboard*, *op. cit.*, Section 7.1.2, p.28.

can be presented per country or city[32], a "hate strength gauge" enabling the representation of hate strength in a specified date range in relation with a specific country. Countries are classified into three categories: non-critical state, state with warning and dangerous state of hate speech usage[33]. Originally, the suspected hate speech rate of a given country was calculated by multiplying the number of suspected hate speech contents originating in this specific city or country with an average hate score of hatefulness (corresponding to the proportion of potentially illegal content to potentially legal ones -as assessed by the content classifier- during the same period of time), divided by the "most hate speech content", which means the total number of hate speech contents assessed as potentially illegal during the same period of time[34]. During research, the hate-rate metric has evolved in order to take into account legal and ethical considerations[35], and the final version of the dashboard shows the number of hateful contents originating from a given country or city taking also into account the total number of contents detected to come from the concerned country during the same period of time. As a consequence, the number of hate speech contents originating in the given city or country multiplied with the average hate score of hatefulness is now divided by the total number of contents (web and tweets) found to originate in this country during the same period of time. Afterwards this value is normalised in order to be displayed as a percentage.

2. A smartphone app

The purpose of the smartphone app is to provide citizens with a user friendly mobile application for easier hate speech reporting, by taking into account user's anonymity. The application will provide:

i) the ability of anonymous reporting of hate-related speech and material found in the web and social media;

ii) compatibility with Android, IOS and Windows mobile devices covering more than 99% of the market (...);

iii) statistical analysis considering hate speech in order to raise the awareness of the user about the impact of hate speech on the world and

iv) a FAQs section[36].

The smartphone app is described in details in the MANDOLA deliverable D3.3 and it does not seem necessary to describe it again so deeply in the current report.

Basically, the smartphone app has been designed so as to enable the user to make a report at a reporting portal. For this purpose, the MANDOLA research consortium provides an API which can be connected with reporting portals which can that way receive the reports. Reports can be done following two different procedures:

- The first one is via the MANDOLA Proxy server, which uses a way back machine[37] and loads the selected URL in a web browser within the application (referred to as

[32] See the MANDOLA Deliverable D3.1 - *MANDOLA Monitoring Dashboard*, *op. cit.*, Section 7.2.4, p.35.

[33] See the MANDOLA Deliverable D3.1 - *MANDOLA Monitoring Dashboard*, *op. cit.*, Section 7.2.9, p.39.

[34] See the MANDOLA Deliverable D3.1 - *MANDOLA Monitoring Dashboard*, *op. cit.*, Section 7.1.2.1, p.29.

[35] Especially the risk for stigmatisation and discrimination, requiring that figures provided by the dashboard are as representative as possible of the situation in relation to hate speech publication in a given country.

[36] See the MANDOLA Deliverable D3.1 - *MANDOLA Monitoring Dashboard*, *op. cit.*, p. 6.

InAppBrowser) in order for the user to highlight the hate speech content. The MANDOLA Proxy is mainly used for public encounters of hate speech in sources such as YouTube, Twitter, news sites and forums, where the content can be viewed publicly and the user is not required to provide any personal information. [38]

- The second method for reporting hate speech is through the OCR - Optical Character Recognition. While the user is browsing with his/her private social media accounts via their native applications, he/she can take a screenshot and report the hate speech encounter without providing any personal information[39]. The processing of the OCR module is done on the device and only the URL of the potentially illegal content is sent in the report (at the exclusion of the screenshot).

The above two methods, OCR and MANDOLA Proxy, have been developed in order to preserve the user's anonymity to the utmost extent and make the reporting easier without reducing the user experience in the social media native applications. Thus, the Smartphone app may run as a background process with the MANDOLA Bubble widget. If activated, this widget listens for a "URL Copy" event and when such an event occurs the widget asks the user if he or she wishes to make a report with this URL. In case of positive answer, the API automatically generates the MANDOLA hate speech report, in order to provide a smoother and easier user experience. The submitted reports are also stored locally in a SQLite database, so as to provide the ability to the user to view and analyse any previous reports. Via the MANDOLA API these reports are also stored into the Report Storage Module, which is an encrypted and secure relational database for keeping the hate speech reports[40].

A hate speech analysis module can also be enabled by the user from the Smartphone app's settings view. It is responsible for determining whether a text provided as input is considered hate-speech and to which categories it can be classified into (e.g. religion, nationality, ethnicity etc.). Where enabled, the possible hate speech content of the report will go through the Hate Speech Analysis module and provide the user with the classifier's output in order to provide him/her with an insight. In order to recognise hate speech, a text classifier was trained with a corpus consisting of annotated hate and non-hate speech text. The corpus used is the one collected and enriched from the social scientists referred to in the presentation of the monitoring dashboard. The corpus is in several languages from the Member States and thus the classifier is language agnostic and performs multi-lingual classification. Furthermore, considering that a potential hate-related input can be labelled with different hate categories, such as religion, nationality and ethnicity, the classifier also supports multi-label

[37] See "Way back machine" on Wikipedia (https://en.wikipedia.org/wiki/Wayback_Machine): "*The Wayback Machine is a digital archive of the World Wide Web and other information on the Internet created by the Internet Archive, a nonprofit organization, based in San Francisco, California, United States.* [...] *The service enables users to see archived versions of web pages across time, which the archive calls a "three dimensional index". Since 1996, the Wayback Machine has been archiving cached pages of websites onto its large cluster of Linux nodes. It revisits sites every few weeks or months and archives a new version. Sites can also be captured on the fly by visitors who enter the site's URL into a search box. The intent is to capture and archive content that otherwise would be lost whenever a site is changed or closed down. The overall vision of the machine's creators is to archive the entire Internet".*

[38] MANDOLA Deliverable D3.1 - *MANDOLA Monitoring Dashboard*, *op. cit.*, p. 7

[39] MANDOLA Deliverable D3.1 - *MANDOLA Monitoring Dashboard*, *op. cit.*, p. 7

[40] MANDOLA Deliverable D3.1 - *MANDOLA Monitoring Dashboard*, *op. cit.*, p. 7.

classification, with the hate categories being the labels. Details can be found in Deliverable D3.1[41].

Beside the above mentioned functions, an awareness module provides the user with statistics in order to raise his or her awareness on online hate speech[42]. It is a scrollable interface where multiple charts, facts and statistics are displayed. This view can also be connected with the MANDOLA Monitoring Dashboard in order to provide the users with dynamic statistical analysis on hate speech. There is also the FAQs view, where the questions and answers from the MANDOLA Deliverable D.4.1 are presented in a more compressed form. The user will be able to view each question's answer within an accordion like component, and also search a question based on keywords[43].

Technically, all the modules are stored on the smartphone, under two reserves:

- The hate speech module will remain on the MANDOLA server at UCY.
- Reports will be received by the third parties connected to the API (such as hotlines fighting against distribution of hate speech), to which the users will send their reports. These reports will be stored in the "report storage module", on the third parties servers.

As regards the degree of finalisation of the smartphone app will be delivered at TRL (Technology Readiness Level) 7: System prototype demonstration in an operational environment. This means that further developments (including testing) are required before commercialisation.

As regards information flows, all the data are stored and analysed on the smartphone, under the following exceptions:

- The hate speech module is stored on the MANDOLA server but does not collect any data. Texts sent by the API are analysed on the fly and classified using a "SGD classifier" and the result is returned to the device. Texts are deleted after analysis.
- Reports will be received by the third parties (connected to the API) to which the users will send their reports, and stored in the "report storage module". Third parties will receive the following data: the identification of the post (URL), the optional title of the report, the date, the category of hate speech, the device ID, and the hate speech containing text (technically, more data could be stored). The user ID will be used in order to warn the user about the action taken on his or her report.

Users will have the possibility to delete information stored on their device.

As regards security, communications between the device, the MANDOLA server and third parties is or will be encrypted using https.

3. A reporting portal

The purpose of the reporting portal is to enable citizens (1) to get information relating to the online hate-speech phenomenon and (2) to find where to report a potentially illegal content that would have been found online.

[41] MANDOLA Deliverable D3.1 - *MANDOLA Monitoring Dashboard*, *op. cit.*, p. 31.

[42] MANDOLA Deliverable D3.1 - *MANDOLA Monitoring Dashboard*, *op. cit.*, p. 7.

[43] MANDOLA Deliverable D3.1 - *MANDOLA Monitoring Dashboard*, *op. cit.*, p. 10.

The reporting portal is described in details in the MANDOLA deliverable D3.2 and it does not seem necessary to describe it again so deeply in the current report.

Basically the information that is provided to Internet users is the following:

- Basic information about the MANDOLA project,
- Existing legal framework related to hate speech and hate crime issues. In this section The MANDOLA research consortium tried to provide in an easy and friendly manner the information that has been produced under the MANDOLA Deliverable D2.1 - Definition of illegal hatred and implication.
- A FAQ section, where visitors of the portal can find answers to the most usual questions related to hate speech, using the MANDOLA deliverable D4.1 - FAQ on Responding to on-line hate speech.

Regarding the possibility to report a potentially illegal online content, the reporting portal only provides links to existing hotlines in each country.

As regards information flows, no Internet users' data is collected by the portal.

As regards security, the reporting portal is hosted by FORTH on their premises in Heraklion. The hosting server is protected by firewalls and is internally and externally monitored in order to minimize the risk from cyber-threats. Additionally, remote backups through the *rsync* utility are performed on a daily basis[44]. The server moreover resides in a protected physical environment. It is located in one of FORTH's data-centres. For ensuring optimal operating environment, it is equipped with industrial-strength air conditioning with more than 240.000BTUs efficiency. In power emergencies, it is supported by a UPS power supply and an external power generator which is engaged automatically on power failure. Additionally, the data-centre features an automatic carbon dioxide fire-extinguishing system[45].

4. Information dedicated to different stakeholders

4.1 policy makers and the Internet Industry

Information dedicated to policy makers aims to provide them with actionable information that can be used to promote policies for mitigating the spread of online hate speech. Information dedicated to the Internet service industry aims to inform these stakeholders about good practices and challenges. This information will be available in the following MANDOLA Deliverables:

- D2.1: Intermediate Report - Definition of Illegal Hatred and Implications
- D2.2: Identification and analysis of the legal and ethical framework
- D2.4b: Privacy Impact Assessment of the MANDOLA outcomes
- D4.4 - Landscape of current responses to hate speech across Europe and gap analysis to avoid duplication of efforts
- D4.2: Best Practice Guide for Responding to Online Hate Speech for Internet Industry and

[44] MANDOLA Deliverable D3.2 - *Reporting Portal*, October 2016, MANDOLA project (Monitoring ANd Detecting OnLine hAte speech) - GA n° JUST/2014/RRAC/AG/HATE/6652, http://mandola-project.eu/, p. 15.

[45] MANDOLA Deliverable D3.2 - *Reporting Portal*, *op. cit.*, p. 16.

- D4.5 - Survey of Industry and LEA to identify best practices and key national exchanges in responding to hate speech online.

4.2 Information dedicated to Internet users

Information dedicated to Internet users aims to assist them in understanding what is hate speech and how to behave when they encounter hate speech online, including what to report, where and how. This information will be available in the following MANDOLA Deliverables:

- D2.1: Intermediate Report - Definition of Illegal Hatred and Implications
- D2.2: Identification and analysis of the legal and ethical framework
- D4.1: FAQ on Responding to on-line hate speech.

A large part of these deliverables will moreover be available from the MANDOLA reporting portal and in the MANDOLA smartphone app.

3.3.1.4 Description of the scope and boundaries of the study

The scope of the study covers the four MANDOLA outcomes described in the previous section, their possible end-users and operators, the different hosting possibilities that may be chosen for their broadcasting and use, and their impact in case their purpose would be diverted. Indeed, the MANDOLA outcomes as they have been conceived appear at first glance to have a very limited impact on rights and freedoms, but -at least as regards technical developments- to be likely to have an important impact in case they would be used for other purposes, eventually in connection with additional technical features.

The scope of the study moreover excludes the situation were the data set used in order to train the hate speech classifier would not be the one provided by UCY or another research organisation, but would be provided and operated by a law enforcement or an intelligence organisation, as well as situations where the latter public stakeholders would operate the monitoring dashboard scanning and collect at this occasion more information that the non-personal information collected by the MANDOLA prototype. Indeed, such kind of data processing would require a particular PIA before implementation, taking into account the particular features of the system, its precise aim and its context of operation. Such a study would go far beyond the limits of the MANDOLA project.

The purpose of this study is the protection of private life, of personal data and of other fundamental rights and freedoms within the framework of the use of the MANDOLA products. Therefore, other end-users activities, outside the implementation, broadcasting and/or use of these outcomes, are not considered.

Challenges at stake are the protection of private life, personal data and other fundamental rights and freedoms, including the social acceptability of the use of the MANDOLA products, while improving the combat against online hate speech.

Participants involved in the study are all the MANDOLA partners, as well as the members of the MANDOLA Advisory Board who have been asked to provide their opinion on the first results of this PIA.

Since the primary challenge at stake is the preservation of private life and personal data, the security operating mode where using MANDOLA technical outcomes is expected to be a "multi-level" mode, where data that might be of a personal nature are collected. In such a

mode, people who are authorised to access the system are not all accredited to the highest classification level, since they do not have all the same need to access the different categories of information handled within the system. This question will be refined during the identification of risks.

3.3.1.5 Identification of the parameters to be considered

Constraints that need to be taken into account cannot all be determined in the current study, since such an identification implies to know the specificities relating to the organisations that will host the systems or some of their components, or that will use the information provided by the MANDOLA consortium (constraints relating to the staff, to the calendar, to the environment, to security requirements, financial and technical constraints... and so forth).

However, the following constraints can be determined. They relate to privacy and personal data protection legal requirements at the Council of Europe and at the European Union level, as the latter has been defined in the MANDOLA deliverable D2.2 (future PIAs could enable the ability to identify other constraints, in accordance with the future E.U. legal framework).

3.3.1.5.1 Privacy and personal data protection legal requirements at the Council of Europe level

The MANDOLA outcomes must firstly comply with the European Convention on Human Rights's (ECHR) requirements. This means that the following principles, which have been identified in Deliverable D2.2[46], must be respected:

- **Legal basis**

  There is currently, at the European or International level, no clear, accessible and foreseeable specific legal basis justifying the collection and/or analysis of online content or of information on a smartphone. Current data protection legislations, as well as the new E.U. General Data Protection Regulation and future domestic legislations implementing the E.U. Directive on personal data protection for the police and criminal justice sector, might be taken as a legal basis if the interference consists of a personal data processing that fully falls within the scope of this legislation, implying *inter alia* that the privacy limitation remains appropriately limited[47], as the MANDOLA consortium conceived the monitoring dashboard (which does not collect personal information beyond those that lie accidentally in the texts that are collected) and the smartphone app (in which the sending of personal information remains under the user's control). However, regular PIA will have to be conducted and any adjunction of technical functions might call for the implementation of additional safeguards, and even to the adoption of a specific legal basis, particularly if the system is intended to be used by public authorities or law enforcement services.

  The legal basis for the outcomes that aim to provide information can be found on Article 10 of the ECHR providing for freedom of expression.

[46] Please refer to the MANDOLA deliverable D2.2 - *Identification and analysis of the legal and ethical framework*, version 2.2.4 of 12 July 2017, MANDOLA project (Monitoring ANd Detecting OnLine hAte speech) - GA n° JUST/2014/RRAC/AG/HATE/6652, http://mandola-project.eu/, for the further description of each of these principles.

[47] See the MANDOLA deliverable D2.2 - *Identification and analysis of the legal and ethical framework*, version 2.2.4 of 12 July 2017, *op. cit*., Section 4.2.3.3.1.

- **Legitimate purpose**

The purpose of the MANDOLA system is to contribute to the combat against online hate speech. In this regards, the purpose is legitimate according to the provisions of the European Convention on Human Rights (ECHR) that protects fundamental rights at stake, including articles 8, 10 and 11 of the ECHR.

- **Necessity**

Regarding the question of whether the MANDOLA outcomes are "necessary", as required by the ECHR, in the light of the questions to be answered in this regard according to the Article 29 Data Protection Working Party[48], we can note the following:

- The MANDOLA outcomes are seeking to address an issue which, if left unaddressed, may result in harm to or have some detrimental effect on society or a section of society. Indeed, a lot of information is already available in relation to hate speech, as well as several initiatives and reporting mechanisms. However, it is still very difficult to understand what is exactly online hate speech, which kind of hate speech is illegal and which actions against hate speech appear to be the most appropriate. This confusion affects both the combat against hate speech (that remains difficult) as well as the protection of fundamental rights (which might be impacted by disproportionate or non-appropriate actions against hate speech).
- There is at this stage no evidence that the MANDOLA outcomes will improve the combat against online hate speech, particularly if the MANDOLA results are ignored, not used or not taken into account. However, the MANDOLA consortium does believe to have created two very interesting instruments, namely a monitoring system and a mobile reporting system, which both preserve to the utmost extent internet users' personal information while making it easier to understand what kind of contents must be reported and to report these contents. In the same line, the MANDOLA consortium have made its best effort to provide policy makers, the Industry and Internet users with a valuable information, as objective as possible, in order to assist them in targeting hate speech more efficiently while avoiding adverse impacts on fundamental rights at stake.
- Existing measures against online hate speech include penal offences, reporting mechanisms, governmental initiatives, private initiative of technical nature, victim assistance, and awareness or information pages and portals. These measures are somewhat successful as shown in the MANDOLA Deliverable D4.4[49]. However, as already outlined above, online hate speech and the understanding of this

[48] Article 29 Data Protection Working Party in its Opinion 01/2014 on the application of necessity and proportionality concepts and data protection within the law enforcement sector (WP 211). See the MANDOLA Deliverable D2.2 Section 4.1.3.2, n°3. See also, British Institute of Human Rights, *Mapping study on projects against hate speech online*, Council of Europe editing 2012, p. 9, 2.1.1, 2, https://rm.coe.int/16807023b4 (last accessed on 25 July 2017): "*The boundaries of what is regarded as hate speech under* [the definition of hate speech proposed by the Committee of Ministers in its Appendix to Recommendation No. R (97) 20 of on "Hate Speech", and retained as a basis of work by several entities] [...] *are likely to fall outside the boundaries of speech which is criminalised under national legislation. They are also likely to fall outside the boundaries of speech which should not be restricted under freedom of expression* [...]. *These are important points because the most common strategy of organisations working in this area appears to be to campaign for greater restrictions on content, or to campaign for content to be taken offline*".

[49] MANDOLA Deliverable D4.4 - *Landscape and Gap Analysis*, August 2017, MANDOLA project (Monitoring ANd Detecting OnLine hAte speech) - GA n° JUST/2014/RRAC/AG/HATE/6652, http://mandola-project.eu/.

phenomenon are still an issue[50], as well as undue limitation of fundamental rights at the occasion of the combat against it[51].The objective and added value of the MANDOLA outcomes is to favour a best identification of potential illegal speeches, a quicker report of these speeches, a better understanding of the phenomenon and to favour best practices in terms of private initiatives, among those that are respectful for other fundamental rights at stake. This last issue is of importance since online hate speech seems to be accompanied with some actions belonging to private justice[52], which constitute a threat for several fundamental rights and freedoms[53].

- In relation to hate speech, opinions that favour criminalisation[54] and preventive actions taken by Internet stakeholders to prevent and delete illegal hate speech on internet servers[55] coexist with opinions that favour de-criminalisation[56], education and adapted public policies[57], and/or call to punish crimes rather than hide them[58]. In addition, several rights such as the rights to privacy and to free speech might be directly impacted by unfettered proactive monitoring[59]. As a consequence, it is of utmost importance, during the PIA of the MANDOLA outcomes, to make sure that technical instruments developed in order to combat hate speech remains confined to the freedoms' limitation that are strictly necessary to pursue the MANDOLA objectives,

[50] MANDOLA Deliverable D4.2 - *Best Practice Guide for Responding to Online Hate Speech for Internet Industry*, March 2017, MANDOLA project (Monitoring ANd Detecting OnLine hAte speech) - GA n° JUST/2014/RRAC/AG/HATE/6652, http://mandola-project.eu/, especially Section 3.4.

[51] MANDOLA Deliverable D4.2 - *Best Practice Guide for Responding to Online Hate Speech for Internet Industry, op. cit.*

[52] MANDOLA Deliverable D4.2 - *Best Practice Guide for Responding to Online Hate Speech for Internet Industry, op. cit.,* p. 8.; Young People Combating Hate Speech On-line, *Mapping study on projects against hate speech online*, prepared by the British Institute of Human Rights, 15 April 2012, Council of Europe publishing 2012 (DDCP-YD/CHS (2012), https://rm.coe.int/16807023b4 (last accessed on 21 August 2017), Section 2.1.1, 2, p. 9.; MANDOLA Deliverable D2.1b - *Definition of Illegal Hatred and Implications, op. cit.,* Sections 5.4 and 6.2.

[53] See for ex. Council of Europe, *Recommendation CM/Rec(2016)5 of the Committee of Ministers to member States on Internet freedom*, https://search.coe.int/cm/Pages/result_details.aspx?ObjectId=09000016806415fa

[54] See for example Peter Moore, *Half of Democrats support a ban on hate speech*, 20 May 2015, https://today.yougov.com/news/2015/05/20/hate-speech/ (last accessed on 25 July 2017).

[55] See for example Eric Silverberg, Carrie Charpentier, Adam Goldman, Karina Luevano, and Jeffrey Petit, *ISP Censorship*, especially in "History of ISP Censorship", https://cs.stanford.edu/people/eroberts/cs181/projects/1998-99/nuremberg-files/censorship.html; Joe Mc Namee, "Self-regulation of content by the online industry", in *The Online Media Self-Regulation Guidebook*, Ed. by A. Hulin and M. Stone, OSCE Representative on Freedom of the Media, 2013, pp. 44 et seq., http://www.osce.org/fom/99560?download=true (URLs last accessed on 25 July 2017).

[56] See for example Sandy Starr, "Understanding Hate Speech"; *in Hate Speech on the Internet*, pp. 10 *et seq.*, http://www.osce.org/fom/13846?download=true; Eric Heinze, *The case against hate speech bans*, 9 April 2014, http://www.eurozine.com/the-case-against-hate-speech-bans/ (also published in *Nineteen arguments for hate speech bans – and against them*, http://freespeechdebate.com/discuss/nineteen-arguments-for-hate-speech-bans-and-against-them/; In defence of hate speech, 15 December 2016, https://www.economist.com/news/leaders/21711914-criminalising-offensive-language-only-empowers-bigots-defence-hate-speech; Peter Tatchell, *Argument – Should hate speech be a crime?*, 1st December 2012, https://newint.org/sections/argument/2012/12/01/is-hate-speech-crime-argument (URLs last accessed on 25 July 2017).

[57] See for example Iginio Gagliardone, Danit Gal, Thiago Alves, Gabriela Martinez, Countering online hate speech, UNESCO, 2015, especially pp. 46 and s., http://unesdoc.unesco.org/images/0023/002332/233231e.pdf.

[58] See European Digital Rights, "Internet blocking - crimes should be punished and not hidden", https://edri.org/wpcontent/uploads/2013/12/blocking_booklet.pdf.

[59] See for ex. Council of Europe, *Recommendation CM/Rec(2016)5 of the Committee of Ministers to member States on Internet freedom, op. cit.;* MANDOLA Deliverable D4.2 - *Best Practice Guide for Responding to Online Hate Speech for Internet Industry, op. cit.,* pp. 29-33, especially pp. 31-32; MANDOLA Deliverable D2.1b - *Definition of Illegal Hatred and Implications*, September 2017, MANDOLA project (Monitoring ANd Detecting OnLine hAte speech) - GA n° JUST/2014/RRAC/AG/HATE/6652, http://mandola-project.eu/.

taking into account this whole context, and that recommendations and advices that are provided take account of all these schools of thought. That way, the MANDOLA outcomes have a chance to be truly necessary since they should bring constructive information to the current debates.

- The MANDOLA research consortium has done everything within its means to take into account any opposition or issue expressed by society, through a study of the legal context and the interview of the Advisory Board members. Privacy by design has been ensured each time it was possible and compatible with the scope of the MANDOLA research, and safeguards that have not been implemented will be the subject of recommendations to end-users and future developers, in order to wind up the PIA.

To conclude, it seems that the MANDOLA outcomes are acceptable, on the "necessity" point of view, provided that they success to bring clarity to the current context, if they fully take into account divergent views and if they provide for sufficient safeguards to make technical developments as less intrusive as possible into people's freedoms. This latter requirement will need to be further assessed in relation to further uses and further technical developments that might be added to the system.

In addition, to be acceptable, the MANDOLA outcomes must be of a proportionate nature.

- **Proportionality**

Regarding the question of whether the MANDOLA outcomes are "proportionate" as required by the ECHR, in the light of the questions to be answered in this regard according to the Article 29 Data Protection Working Party[60] (which amounts to the question of whether they are strictly necessary and surrounded by appropriate safeguards), we can note that they indeed may be considered as "strictly necessary" in relation to their context, to their scope and to their nature, provided that a series of identified safeguards are in place and respected, including MANDOLA recommendations of use and further development.

Strictly necessary in relation to their context

- The MANDOLA outcomes shed light on the hate speech definition, context and issues; help the assessment of its spread online; answer Internet users' questions and ease their reports (of which only a minimum amount of information is kept). These actions seem to be adapted to the severity of the social need, which is to clarify the context, the phenomenon, and to identify the appropriate means to combat hate while considering all the opinions in this field and fundamental rights preservation requirements.

  However, these outcomes will be appropriate as long as they remain up-to-date. As a consequence, regular PIAs of the technical systems must inter alia ensure that statistics stay up-to-date and consider potential legislative changes, and that the reporting systems stay user-friendly despite modifications of the technical

[60] Article 29 Data Protection Working Party in its Opinion 01/2014 on the application of necessity and proportionality concepts and data protection within the law enforcement sector (WP 211). See the MANDOLA deliverable D2.2 - *Identification and analysis of the legal and ethical framework*, version 2.2.4 of 12 July 2017, MANDOLA project (Monitoring ANd Detecting OnLine hAte speech) - GA n° JUST/2014/RRAC/AG/HATE/6652, http://mandola-project.eu/, Section 4.1.3.2, 4.

environment. In the same line, regular PIAs (which might be small-scale ones[61]) must be performed in relation to the broadcasting and use of the MANDOLA information provided to policy makers, to the industry and to Internet users, particularly where this information is susceptible to be obsolete or if assets of sciences that have been considered in order to write this information are susceptible to have evolved. This, unless a disclaimer sheds clearly light on the information's date of production, and warns on a risk of obsolescence after a certain period of time.

- The behaviour that is intended to be restricted (through the combat against online hate speech) is to hurt people in their dignity and right to non-discrimination, which is not a legitimate behaviour. Moreover, the MANDOLA outcomes are adapted to the new challenge which is the increasing use of the Internet in order to spread hate speech[62]. All these elements show that the MANDOLA outcomes seem to be adapted to their context.

Strictly necessary in relation to their scope

- **In relation to the volume of information collected**, the MANDOLA technical outcomes do not collect direct personal data, and very few indirect ones: some texts (which might accidentally contain identification signs) and URLs (which might lead to specific contents, and therefore to their author) might be collected by third parties connected to the smartphone app, under their responsibility, either to train the MANDOLA hate speech classifier, or to handle hate speech reports (which are stored in the report storage module). These third parties do also receive the device ID of the persons making a report in order to inform these persons about the action taken on their report, but the sending of this ID can be desactivated by the user of the smartphone app as a privacy-by-design safeguard. Therefore, if the purposes and restrictions of these technical mechanisms are not diverted, no personal data are processed by these systems, unless otherwise agreed and chosen by the user, or accidentally, because of the nature and content of the sources that are scanned or because it is the only way to inform back the author of a report. In any case, identifying people (either authors of reports or authors of potentially illegal content) is not an objective of the system.

  Consequently, since the system has no interest in individuals, the interference with private life is limited compared to systems that aim at listing individuals with other kind of data linked to them. However, since personal data might be processed accidentally or through a diversion of its purposes and restrictions, data protection remains important and safeguards that need to be implemented include compliance with personal data protection legislation, with special care dedicated to the non-diversion of the tool without further legal and ethical analysis and assessment.

  In addition, third parties connected to the system might collect personal data, as already explained, such as the device ID of the author of a given report (with his or her

[61] See the MANDOLA Deliverable D2.4a (Intermediate) - *Privacy Impact Assessment of the MANDOLA outcomes*, version 2.4a.2 of 11 July 2017, MANDOLA project (Monitoring And Detecting Online hate speech) - GA n° JUST/2014/RRAC/AG/HATE/6652, http://mandola-project.eu/, Section 4.1.

[62] See for example Young People Combating Hate Speech On-line, Mapping study on projects against hate speech online, prepared by the British Institute of Human Rights, 15 April 2012, Council of Europe publishing 2012 (DDCP-YD/CHS (2012), https://rm.coe.int/16807023b4 (last accessed on 21 August 2017), Section 4, p. 15; Steve Lahr, Online Hate Sites Grow With Social Networks, 16 March 2010, https://bits.blogs.nytimes.com/2010/03/16/online-hate-sites-grow-with-social-networks/ (last accessed on 25 July 2017).

agreement, as a safeguard), such as additional comments sent voluntarily to these third parties, and such as the texts and URLs of certain potentially illegal contents, which might inter alia contain personal data relating to the victim of the speech[63]. In this regard, some safeguards should ideally be implemented. The first one would be the obligation made to each third party willing to be connected to the system to ensure a full compliance of its personal data processing with law, including short time-limitation and security of processing. A second one should be the possibility, offered to the end-user of the smartphone app, (1) to visualise, prior the sending of his or her report, the name of the recipients proposed for his or her particular report and the detailed personal data policy of this or these third party(ies), and (2) to remove one or several of these recipients and to choose new one(s), eventually from a predefined list. Indeed, the app is currently configured in a way that the report is sent automatically to the relevant assistant service connected to the app. As a consequence, Internet users who want to report hate speech through the app while wishing to receive a feedback on the action taken on this report are not totally free to choose who will process their personal data.

- **In relation to the information provided to Internet users, policy makers and the industry** (through the smartphone app, the reporting portal and the MANDOLA website), the extent of freedoms' limitation will depend on the content of this information, and on the way its recipients will use it. It will be therefore important to provide information as objective and exhaustive as possible, with appropriate disclaimers where particular information might lead to behaviours infringing fundamental rights.
- **In relation to the additional information provided by the monitoring dashboard**, a particular issue might concern the hate map[64], which must provide exact information in order to not mislead viewers as to the nature of the statistics that are shown. Indeed, the monitoring dashboard aims at showing potentially illegal hate speech, as the notion has been defined in the MANDOLA deliverable D2.1 and as such contents have been assessed technically, by a tool trained humanly by some social analysts. However, the latter do not have the skills and the independency of judges from the judiciary, and have received some texts to assess, which have been taken out of their original contexts. In these conditions, and in a context where (1) a given content will be illegal only if a judge declares this illegality after having established all the elements -including of context- that are part of the penal offence, and where (2) legislations on hate speech are very complex and differ widely between countries[65], the dashboard results must be handled with care. To this end, the dashboard results must be accompanied by a disclaimer explaining the above-mentioned context. In addition, during the subsequent development phases of the dashboard, some research could ideally focus on ways to improve the accuracy of results (while testing this accuracy), for example by applying different criteria of content assessment taking into account

[63] This clarification relating to victims' personal data has been added following consultation of the Mandola Advisory Board members.

[64] See the MANDOLA Deliverable D3.1 - *MANDOLA Monitoring Dashboard*, MANDOLA project (Monitoring And Detecting OnLine hAte speech) - GA n° JUST/2014/RRAC/AG/HATE/6652, http://mandola-project.eu/, p. 24.

[65] MANDOLA Deliverable D2.1b - *Definition of Illegal Hatred and Implications*, September 2017, MANDOLA project (Monitoring ANd Detecting OnLine hAte speech) - GA n° JUST/2014/RRAC/AG/HATE/6652, http://mandola-project.eu/.

the probable competent jurisdiction, and by finding ways to take the context of the speech into account during this assessment (such as cultural aspects[66], the author's intent[67], polarity[68], or the existence of a public disorder[69] based on the relevant country's courts decisions).

In addition to add a disclaimer to general results and to the classification of content according to the category they belong to, disclaimers as safeguards appear to be particularly important in relation to:

- ✓ the percentage of hate-speech shown by country and city[70], since such results could lead to the stigmatisation and even to the discrimination of a whole country or city and therefore of people belonging to these countries and cities, whereas the data used to obtain these results (1) can be wrong or partial, (2) take into account contents that might be not illegal (since illegality criteria vary from country to country) and (3) are not calculated taking into account certain parameters such as the difference between countries in terms of internet penetration and of number of Internet users[71], as well as hazards that might lead some persons to express themselves at certain period of time and not during other periods. As a consequence, results might lead to give a wrong idea of the situation, especially in terms of percentage per inhabitant in capacity to produce a hate speech content.
- ✓ the Hate strength gauge[72] which enables to obtain a gauge representation of hate strength in specified date range and country. Based on the percentage of hate strength, it enables to classify countries into three categories: non-critical / warning / dangerous state of hate speech usage[73]. This might also lead to (false) stigmatisation and discrimination of the country and of its inhabitants, keeping in mind that (1) this classification might be meaningless if it does not take into account some other parameters such as the volume of the population, the Internet penetration and the number of Internet users compared to other countries, and that (2) even if these latter factors were taken into account, a high level of online hate speeches does not necessarily means that a given country as a whole is dangerous in terms of hate speech usage. Consequently, a MANDOLA

[66] For example, in certain regions, hate speech can be culturally trivialised without intent of inciting hate (cultural aspects and the current footnote have been added following consultation of the Mandola Advisory Board members).

[67] Intention is of high importance and should always be one of the constitutive elements of a hate speech offence. For example, hate speech can be used in the text of a theatre piece in the purpose of denouncing hate. During the MANDOLA Advisory Board consultation, it has been emphasised that hate-speech words can be used for other purposes than hate speech.

[68] In the extension of the previous footnote, one member of the MANDOLA Advisory Board emphasised that hate speech can exist through metaphors and words shared by some people only.

[69] Which might be a requirement, in certain jurisdiction, in order to consider illegal a hate content. For further details see the MANDOLA deliverable 2.1 - *Definition of Illegal Hatred and Implications*, September 2017, MANDOLA project (Monitoring ANd Detecting OnLine hAte speech) - GA n° JUST/2014/RRAC/AG/HATE/6652, http://mandola-project.eu/.

[70] See the MANDOLA Deliverable D3.1 - *MANDOLA Monitoring Dashboard, op. cit.*, p.29; pp.35-38.

[71] This clarification has been added following the consultation of the MANDOLA Advisory Board.

[72] See the MANDOLA Deliverable D3.1 - *MANDOLA Monitoring Dashboard, op. cit.*, p. 39.

[73] As it has been specified in Deliverable D3.1, *op. cit*.

Advisory Board member warned on the necessity to non-use the term "dangerous" in order to quality the state of hate speech usage[74].

- As a result, a disclaimer (visible where the results per country or per city and the results of the Hate strength gauge are displayed) should detail very clearly the variables that are taken into account in order to calculate the hate speech score of countries and cities (such as the number of inhabitants and the volume of Internet content produced each day), avoid the use of the word "dangerous" and explain on the opposite in simple terms that these statistics cannot represent the state of dangerousness of a given country or city, in particular since (1) they don't take into account several important factors such as the Internet penetration, the number of Internet users and the frequency and their habits in terms of Internet usage; (2) the MANDOLA dashboard shows hate speech that is potentially illegal in one or several E.U. countries but that might not be illegal in one or several others; (3) the context of the speeches are not taken into account and the assessment of contents is not exact science; and (4) even a high level of illegal online hate speeches (which might be produced by the same group of persons, and which are eased by the simplicity of posting on the Internet) does not necessarily means that a given country or city as a whole is dangerous in terms of hate speech usage.
- In addition, in relation with both these categories of results, further research should aim at also presenting results that would take into account the most possible relevant factors such as the number of inhabitants, the Internet penetration, the number of Internet users and the frequency and their habits in terms of Internet usage.
- **In relation with additional information provided through the smartphone app**, similar issues do concern:
  - ✓ The provision of statistics coming from the dashboard, which should include the same safeguards as referred to above;
  - ✓ The provision of a FAQ, which should include the same safeguards as referred to two points above;
  - ✓ The possibility to send reports from private areas and to analyse them in different languages that will be previously downloaded on the smartphone. Indeed, private areas might contain private hate speech, which is mainly not considered to be illegal in the studied E.U. countries. In the same line, a content that might be illegal or perceived as illegal in one given country might be legal in one other, which means that contents written in different languages should be ideally assessed differently. As a result, in order to make the context perfectly clear to the smartphone app user, disclaimers should be added at the level of both these functions in order to shed light on their context and on the precautions to be taken in this regard.
- **Persons concerned by the processing operations that feed the monitoring dashboard and the smartphone app** might be numerous, since the first one scans the web and social networks, while the second one might be downloaded by any Internet user possessing a smartphone running the Android or Windows iOS. However, as already mentioned, the monitoring dashboard does not collect personal data (beyond some

[74] Within the framework of the Advisory Board consultation (see Section 3.6 of the current report)

texts and URLs that might be stored separately by third parties in order to train the system, and that might contain personal data accidentally) and the smartphone app is supposed to only collect data sent voluntarily by the reporting person (his or her consent being currently impaired by the impossibility to choose the recipients of the report, which could be solved by developing the possibility of such choice in the future). In this context, the number of people concerned does not seem to be an issue as long as the MANDOLA technical outcomes are not further developed in order to collect additional personal data and are not used on other purposes, and if recommended safeguards are implemented (disclaimers, user's choice in relation with the recipients of his report as well as in relation with the sending of his or her device ID, and legal compliance, including of third parties processing).

- **Persons concerned by the MANDOLA outcomes consisting of providing information** might also be numerous, depending on the MANDOLA outcomes visibility. This argues in favour of taking a particular care in the implementation of the safeguards referred to previously in this discussion (namely to provide an information as objective, exhaustive and referenced as possible, with appropriate disclaimers where particular information might lead to behaviours infringing fundamental rights).
- Finally, on the question of whether the MANDOLA outcomes leave some scope for the potentially limited freedoms[75], including privacy, the answer is clearly yes, taking into account the elements analysed above and providing that the purposes of technical mechanisms and their restrictions of use (implemented or recommended by the MANDOLA consortium) are respected. These last points will be crucial since the identification of individuals whose data are processed may lead to the possibility to either monitor the behaviour of a really high number of natural persons, potentially for any purpose, especially in order to search for the commission of penal infringements (while, in this case, no evidence of a criminal behaviour would have been collected prior to the setting up of such monitoring, and before any crime would have been qualified[76]). This could not take place before the performance of a specific PIA identifying the possibility to pursue such aim using such systems, and, if so, identifying needed appropriate safeguards.

Strictly necessary in relation to their nature

- The last question regarding proportionality is to know if other measures, of a less intrusive nature, could be considered, and, if yes, why they have been rejected.

  The answer seems to be negative in relation to the provision of information, since basic information does already exist in the field of the combat against hate speech and since the benefits of the MANDOLA project are expected to partly lie in the broadness of its legal and field research.

  The question is more difficult to answer at the MANDOLA research level in relation to the monitoring dashboard and the smartphone app. Indeed, the content of the project has been defined before it started, and its purpose was not to investigate the

[75] In other words, if they do not extinguish the possibility to exercise these freedoms. See the MANDOLA Deliverable D2.2 - Identification and analysis of the legal and ethical framework, version 2.2.4 of 12 July 2017, MANDOLA project (Monitoring ANd Detecting OnLine hAte speech) - GA n° JUST/2014/RRAC/AG/HATE/6652, http://mandola-project.eu/, Section 4.1.3.

[76] This would be contrary to the ECtHR court cases and most national constitutional laws.

possibility to create other systems than the ones that have been promised, but only to ensure the relevance of these systems and to investigate the possibility of building them in the best possible way.

However, it appears that producing statistics on online hate speech is a need[77], and that it necessarily involves the scan of web content. It also appears that an easy-to-use reporting system is a victims' need[78], and is particularly beneficial on smartphones, which are widely used today to access the Internet. In addition, only a smartphone app enables to provide the functionalities proposed by the MANDOLA consortium, knowing that smartphone users stay free to use other reporting mechanisms including the MANDOLA reporting portal instead of the app, and to send their report directly to the relevant assistance services, using the web. In this sense, it does not appear that less intrusive techniques of another nature could have been proposed.

Limited by appropriate safeguards (summary)

For clarity reasons, preceding analyses already provided for recommendations relating to the safeguards that appear to be needed in order to palliate weaknesses of the previous steps of the assessment. These recommended safeguards can be summarised as follows:

- Compliance with ECHR requirements of subsequent uses of the MANDOLA outcomes implies to respect the MANDOLA recommendations of use and of further development. Purposes and restriction of use (including implemented safeguards) of the monitoring dashboard and of the smartphone app must especially not be diverted.
- Technical developments must remain confined to the freedoms' limitation that are strictly necessary to the MANDOLA objectives (to favour a best identification of potential illegal speeches, a quicker report of these speeches, a better understanding of the phenomenon and to favour best practices in terms of private initiatives), and recommendations and advices that are provided must take into account all the schools of thought in the combat against hate speech area.
- Implementation and use (for other entities than Internet users) of the monitoring dashboard and of the smartphone app require compliance with personal data protection legislation, with special care dedicated to the non-diversion of the tools without further legal and ethical analysis and assessment. In particular, third parties connected to the smartphone app and collecting personal data (including indirect ones such as devices ID, Internet texts and their URLs) should have the obligation to ensure a full compliance of their personal data processing with law, including short time-limitation and security of processing, as well as providing their personal data policies to be shown to the Internet user;
- Internet users using the smartphone app must have the possibility to not send the device ID of their smartphone to third parties, being warned that, if they make his choice, they will not be informed on the action taken on their report. The MANDOLA consortium has already provided for this functionality which must not be removed.

[77] See for example the European Union Agency for fundamental rights, *Hate crime,* http://fra.europa.eu/en/theme/hate-crime; Ester Strømmen, *Hate Crime and Hate Speech in the European Union,* 6 Nov. 2014, Foreign Affairs Review, http://foreignaffairsreview.co.uk/2014/11/hate-crime/ (URLs last accessed on 12 September 2017).

[78] See for example the European Union Agency for fundamental rights, *Hate crime, op. cit.*

- Internet users using the smartphone app should have the possibility (1) to visualise, prior the sending of their report, the name of the recipients proposed for one given report, (2) to visualise the detailed personal data policy of this or these third party(ies), and (3) to remove one or several of these recipients and to choose new one(s), eventually from a predefined list.
- The dashboard results must be accompanied by a disclaimer explaining their context and the care to be taken when interpreting them. This applies to the dashboard general results and to the classification of hate speech into categories, but also, in particular,
  - ✓ To the dashboard results showing hate-speech by country and city[79] , since such results could lead to the stigmatisation and even to the discrimination of the country as a whole, and therefore of people belonging to these countries and cities.
  - ✓ To the Hate strength gauge[80] which enables to obtain a gauge representation of hate strength in specified date range and country. Based on the percentage of hate strength, it enables to classify countries into three categories: non-critical / warning / dangerous state of hate speech usage. This might also lead to (false) stigmatisation and discrimination of the country and of its inhabitants.
- As a result, a disclaimer (visible where the results per country or per city and the results of the Hate strength gauge are displayed) should detail very clearly the variables that are taken into account in order to calculate the hate speech score of countries and cities (such as the number of inhabitants and the volume of Internet content produced each day), avoid the use of the word “dangerous” and explain on the opposite in simple terms that these statistics cannot represent the state of dangerousness of a given country or city, in particular since (1) they don’t take into account several important factors such as the Internet penetration, the number of Internet users and the frequency and their habits in terms of Internet usage; (2) the MANDOLA dashboard shows hate speech that is potentially illegal in one or several E.U. countries but that might not be illegal in one or several others; (3) the context of the speeches are not taken into account and the assessment of contents is not exact science; and (4) even a high level of illegal online hate speeches (which might be produced by the same group of persons, and which are eased by the simplicity of posting on the Internet) does not necessarily means that a given country or city as a whole is dangerous in terms of hate speech usage.
- During the subsequent development phases of the dashboard, some research should focus on ways to improve the accuracy of results (while testing this accuracy), by taking into account the most possible relevant factors such as:
  - ✓ The number of inhabitants, the Internet penetration, the number of Internet users and the frequency and their habits in terms of Internet usage;
  - ✓ The probable competent jurisdiction;

[79] See the MANDOLA Deliverable D3.1 - *MANDOLA Monitoring Dashboard, op. cit.*, p. 29; pp.35-38.

[80] See the MANDOLA Deliverable D3.1 - *MANDOLA Monitoring Dashboard, op. cit.*, p. 39.

- ✓ The context of the speech such as cultural aspects[81], the author's intent[82], polarity[83], or the existence of a public disorder[84] based on the relevant country's courts decisions.

- o In the smartphone app, disclaimers should be added at the level of both the possibility to send reports from private areas and to analyse reports in different languages, in order to shed light on their context and on the precautions to be taken in this regard. Indeed, private areas might contain private hate speech, which is mainly not considered to be illegal in the studied E.U. countries. In the same line, a content that might be illegal or perceived as illegal in one given country might be legal in one other, which means that contents written in different languages should be ideally assessed differently.
- o Information to be provided by the MANDOLA consortium to Internet users (including a FAQ), policy makers and the industry, through the MANDOLA portal, the MANDOLA reporting portal, the smartphone app and the monitoring dashboard, must be as objective, exhaustive and referenced as possible, with appropriate disclaimers where a given information might encourage behaviours infringing fundamental rights. In particular, it must favour best practices in terms of private initiatives, among those that are respectful for other fundamental rights at stake. This last issue is of importance since online hate speech seems to be accompanied with some actions belonging to private justice[85], which constitute a threat for several fundamental rights and freedoms[86].
- o Regular PIAs of the technical outcomes will have to be conducted by data or system controllers[87] and any modification of purposes or adjunction of technical functions imply the performance of a specific PIA in order to identify the possibility to pursue the new purposes or to implement the new functions, and, if so, in order to identify the new appropriate safeguards that are needed in this regard (including a specific legal

[81] For example, in certain regions, hate speech can be culturally trivialised without intent of inciting hate (cultural aspects and the current footnote have been added following consultation of the Mandola Advisory Board members).

[82] Intention is of high importance and should always be one of the constitutive elements of a hate speech offence. For example, hate speech can be used in the text of a theatre piece in the purpose of denouncing hate. During the MANDOLA Advisory Board consultation, it has been emphasised that hate-speech words can be used for other purposes than hate speech.

[83] In the extension of the previous footnote, one member of the MANDOLA Advisory Board emphasised that hate speech can exist through metaphors and words shared by some people only.

[84] Which might be a requirement, in certain jurisdiction, in order to consider illegal a hate content. For further details see the MANDOLA deliverable 2.1 - *Definition of Illegal Hatred and Implications*, September 2017, MANDOLA project (Monitoring ANd Detecting OnLine hAte speech) - GA n° JUST/2014/RRAC/AG/HATE/6652, http://mandola-project.eu/.

[85] See the MANDOLA Deliverable D4.2 - *Best Practice Guide for Responding to Online Hate Speech for Internet Industry,* March 2017, p. 8; Young People Combating Hate Speech On-line, *Mapping study on projects against hate speech online*, prepared by the British Institute of Human Rights, 15 April 2012, Council of Europe publishing 2012 (DDCP-YD/CHS (2012), https://rm.coe.int/16807023b4 (last accessed on 21 August 2017), Section 2.1.1, 2, p. 9.; MANDOLA Deliverable D2.1b - *Definition of Illegal Hatred and Implications, op. cit.,* Sections 5.4 and 6.2.

[86] See for ex. Council of Europe, *Recommendation CM/Rec(2016)5 of the Committee of Ministers to member States on Internet freedom*, https://search.coe.int/cm/Pages/result_details.aspx?ObjectId=09000016806415fa

[87] The Article 29 Data Protection Working Party advises to conduct a review at the latest every three years in its *Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679* (WP248), 4 April 2017, http://ec.europa.eu/newsroom/document.cfm?doc_id=44137 (last accessed on 15 June 2017), p. 12.

basis, particularly if the system is intended to be used by public authorities or law enforcement services). Regular PIAs will *inter alia* need to ensure that statistics stay up-to-date and consider potential legislative changes, and that the reporting systems stay user-friendly despite modifications of the technical environment.

- Regular PIAs (which might be small-scale ones[88]) must be performed in relation to the broadcasting and use of the MANDOLA information provided to policy makers, to the industry and to Internet users, particularly where this information is susceptible to be obsolete or if assets of sciences that have been considered in order to write this information are susceptible to have evolved. This, unless a disclaimer sheds clearly light on the information's date of production, and warns on a risk of obsolescence after a certain period of time.

3.3.1.5.2 European personal data protection legal requirements

In addition to the requirements of the ECHR, the MANDOLA outcomes must comply with data protection legal requirements. This means that the following principles, which have been identified in Deliverable D2.2[89], must be respected in relation with (1) the smartphone app (in relation to (a) the module that might train the hate speech classifier, to (b) the reception and storage of reports and of device IDs by assistance services, including potentially LEAs, and (c) in a relative manner to the data stored and processed by the app on the user device), and (2) the components of the monitoring dashboard that will or might process personal data (namely the module - called ground truth data set - that might train the hate speech classifier and the analysis on the fly of scanned web and social networks contents, which might contain direct or indirect personal data). Other components of the monitoring dashboard, the reporting portal and informational outcomes of the MANDOLA research are not concerned by this legal compliance test, since they are not processing any even indirect personal data.

- **Legal basis**

The MANDOLA technical outcomes do not seem to need an additional legal basis than current legislation on personal data protection, which will be replaced in 2018 by the E.U. General Data Protection Regulation and future domestic legislations implementing the latter as well as the E.U. Directive on personal data protection for the police and criminal justice sector. However this conclusion might be different in case other functions and / or other purposes would be added to these systems.

- **Legitimate, explicit and specified purpose**

The purpose of assisting in the combat against online hate speech by shedding light on the hate speech definition, context and issues, by helping the assessment of its spread online, and by answering Internet users' questions and easing their reports, is legitimate, since it is

[88] See the MANDOLA Deliverable D2.4a (Intermediate) - *Privacy Impact Assessment of the MANDOLA outcomes*, version 2.4a.2 of 11 July 2017, MANDOLA project (Monitoring ANd Detecting OnLine hAte speech) - GA n° JUST/2014/RRAC/AG/HATE/6652, http://mandola-project.eu/, Section 4.1.

[89] For a more detailed description of each of these principles please refer to the MANDOLA deliverable D2.2 - *Identification and analysis of the legal and ethical framework*, version 2.2.4 of 12 July 2017, MANDOLA project (Monitoring ANd Detecting OnLine hAte speech) - GA n° JUST/2014/RRAC/AG/HATE/6652, http://mandola-project.eu/.

in accordance with the law in the broadest sense (under the reserve of the necessity and proportionality of the measure).

Such a purpose is also specified, in other words sufficiently defined prior the time of the data collection, provided that end-users do not use the MANDOLA developments for other purposes, and provided that assistance services that will receive reports through the smartphone app do only process these reports and other related personal data in the solely aim of analysing the potential illegality of contents, before forwarding them to competent authorities and/or organisations, and answering the author of the report.

Processing operations that are and are not included in this specification have been identified and detailed in several MANDOLA deliverables, and are summarised in Section 3.3.1.3 of the current report. This description, as well as the precise purposes of operations, will have to be included in the MANDOLA recommendations of use, along with the purposes and processing operations that are authorised to third parties that will receive reports through the app.

Such a purpose is moreover explicit, and appears to be understandable by anyone. However, the monitoring dashboard and the smartphone app, as well as all the supports and channels that will give access to the MANDOLA dashboard results, will have to include a visible and consistent information in this regard, in order to ensure the predictability of processing operations for data subjects, before processing operations take place.

Regarding the question to know whether the potential collection or processing of personal data by (1) the scanning component of the monitoring dashboard and by (2) the module (the ground truth data set or other similar module) that trains the hate speech classifier (which might operate within the framework of the dashboard and of the smartphone app) are compatible with the first processing that led to the publication of the information on the public Internet[90], we can note the following, in the light of the compatibility test proposed by the Article 29 Data Protection Working Party[91]:

- On the first hand, the intended purpose of the publication of an article on a webpage, or of a comment on such a webpage, for the internet user, is the publicity of the information to the entire internet. The intended purpose of a publication on the public part of a social network, for the Internet user, is the publicity of this information to the other users of this social network, and eventually to all Internet users if the author chooses to be visible from search engines. However, some social networks impose to Internet users to make publicly available certain identified information, such as their name. The purpose of the publicity of such information will be, in these cases, the wish to be present on this social network.

  On the other hand, the intended purpose of the MANDOLA scanning component of the dashboard is to analyse the afore-mentioned information (in case the afore-mentioned information sources are scanned) in order to (1) detect its potential illegal nature, (2) classify it into a Subcategory of illegal contents, (3) collect some non-personal data

[90] We do not consider the collection of data on Internet's private areas, which is not projected within the framework of both the monitoring dashboard and the smartphone app, and which would imply a higher interference with privacy and other rights, and should be covered by national rules relating to private electronic communications / correspondences interception.

[91] Article 29 Data Protection Working Party, Opinion 03/2013 on purpose limitation, 2 April 2013, WP203, III.2.2, p.23 *et seq*.

namely geolocation (timestamp of the tweets where the user has enabled the location service) reduced to 3 decimals in order to avoid any identification of a particular individual.

Regarding the intended purpose of the MANDOLA ground truth data set or other similar module, it is to (1) collect potentially illegal texts that have been afterward or will be analysed by social scientists, to (2) store them in a database separated from the rest of the system, and to (3) train the hate speech classifier thanks to the data they contain[92].

- Regarding the content of the relation between these two groups of purposes, it consists in the possibility given to the public (of one given social network or more generally of the Internet) to access personal information, and, as a consequence, to use this information according to its context of publication. Therefore, potential personal data processing operated by the MANDOLA technical outcomes might partly going beyond this relation, since (1) they might include data coming from persons who would have liked to restrict the publication of these data to the natural persons using the concerned social network only, and since (2) they aim at using information out its context of publication by (2a) extracting from it non-personal information or (2b) extracting from it some data, further analysed by human and non-human means, in order to train a hate classifier. However, in the first case the use of the information cannot impact the data subject since the latter cannot be recognised, and in the second case the indirect personal data that might potentially be included in collected texts are not supposed to be used, neither by the system nor by humans.
- Regarding reasonable expectations of Internet users or of the users of potentially scanned social networks, in terms of private life protection, the possibility that a third party will access and therefore use their information is at least known by users, as soon as they do accept this publicity, either in the case this publicity has been freely chosen or in the case it has been partly imposed by the general conditions of use of the concerned social network (publications that have been partly "imposed" by a social network - usually at least name and photographies - may lead to a higher expectation of privacy, in terms of non-reuse of this information, but are not processed by the MANDOLA technical developments). For the rest, all the Internet users who publish information on the web or on social networks will not all the time be conscious of the existing possibility of analysing deeply this information in order to correlate it with other data, and to deduce from the whole other information. As a consequence, the MANDOLA technical operations might (depending on the sources that are used) go beyond reasonable expectations of Internet users in terms of private life respect, where they relate to information that might be linked to the direct or indirect identity of a given internet user.

[92] The assessment of the illegality of a content is based on a) the classification algorithm and b) the ground truth data set. Specifically, UCY has used the Stochastic Gradient Descent classifier from the following library: http://scikit-learn.org/stable/modules/sgd.html#classification. The SGD Classifier has a number of parameters (see http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html#sklearn.linear_model.SGDClassifier). Using the Randomised search method (see http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html), the research team has searched for the best parameters for the classifiers (based on their data set).

- Regarding the sensitivity of concerned data and the impact of a further processing of these data, especially in emotional terms, scanned data involved in the MANDOLA processing operations might be of all types, due to the nature of the sources that are scanned. Therefore, these data might be sensitive within the framework of the training of the hate speech classifier (no potentially sensitive data are collected outside this component), particularly if social networks are involved, since Internet users use often such networks to publish very personal information (such as philosophical or religious views, health...), and since oral or written hate-filled statements might involve sensitive data relating to victims[93]. Mentions of criminal infringements (linked or not to an individual) may also be collected and processed, since it is highly possible that they appear in searched documents, given the subject of the research which is hate speech. The impact of a further processing might be high, if such a processing drives to take decisions against individuals or to enhance the publicity of the data involved. The existence of the processing itself, if Internet users are informed of the processing but not of the safeguards put in place to protect rights and freedoms, may drive to restrict other freedoms than the right to private life, for example the freedom of expression, the freedom to communicate and the freedom to develop relationships with other human beings, if this information generates self-censorship behaviours.
- Regarding finally the safeguards that should be implemented to compensate for the weaknesses found out during the previous steps of the compatibility evaluation, answers brought to these steps drive to the following conclusions:
  - ✓ Ideally, no individual should be identifiable in the MANDOLA hate database. This implies to remove all names and other visible signs that might lead to or that might be of a personal nature. Since simple texts might occasionally lead to identify an individual, even if all visible personal data are removed, this also implies the application of security measures against undue internal or external access, in order to ensure that access to a text or to a URL stored in the hate speech database pursues the solely aim of verifying the illegal nature of one given content, in order to enhance the performances of the dashboard.
    - This could be ensured by restricting the access to the information contained in the database and to originating URLs to identified persons accredited to do it on a "need to know" or "need to use" basis, and by implementing access control and recording, a regular independent control of these accesses and of their purposes, and agreements of confidentiality and of non-misuse.
    - This could also be ensured by avoiding recourse to hosting providers, and where impossible by strong contractual and security measures that would prevent any undue access, modification, record or other processing of data by a hosting provider or a technical provider which services would be used by the operators of a hate speech database.
    - In addition, a regular deletion of URLs and of all the texts that might contain indirect personal data, as long as they are not required for the proper functioning of the system, should be planned.

[93] This second clarification (relating to victims' personal data) has been added following consultation of the Mandola Advisory Board members.

- ✓ The mechanism that removes a part of the geolocation coordinates in order to anonymise data is of utmost importance and must be particularly preserved and secured against removal or circumvention.
- ✓ The fact that a name of person or a sign/a sentence that might identify a person, included in the database, will never correspond for sure to a real and identifiable natural person, must be very clear for any user of the MANDOLA hate speech database.
- ✓ Strategic decisions taken on the basis of the MANDOLA monitoring dashboard should not restrict some individual's rights. If such a restriction of rights is planned, these decisions should not be taken if not corroborated by additional information obtained from outside the MANDOLA system.
- ✓ Transparency should be ensured regarding the exact nature of potentially personal data that may be included in the database, the purposes of the processing, data sources and measures put in place in order to ensure the protection of privacy and personal data.
- ✓ Measures of awareness raising, control and enforcement must ensure that no modification of the conditions of use of the MANDOLA technical developments (such as they are described in the current report and other MANDOLA technical deliverables) are tolerated, notably at the occasion of the integration of another component to the systems, or/and by using these systems in order to identify individuals or in order to process voluntarily personal data, without performing a new identification of the appropriate safeguards that must be implemented (the best method being the performance of a PIA that would take into account these modifications and would follow again each step of the PIA method proposed in the MANDOLA Deliverable D2.4a), since the conclusions of the current PIA as a whole and of each of its steps would not be adapted anymore to such a new situation.

- **Data quality**
  - Fairness of the processing is and will remain an obligation for all entities, even though the latter statement can be slightly relativised for the police and criminal justice sector (for the latter, it is not an obligation under all legal instruments applicable to their activities but it will become mandatory within the framework of the future E.U. Directive on personal data protection for the police and criminal justice sector). Within the context of the MANDOLA system, the issue of the fairness of the processing appears to be connected to the issue of the compatibility of the further processing, regarding the reasons that lead the Internet user to publish information on the Internet, since the conditions to be verified are in both cases the same. This question has already been analysed in the previous point of the current study.
  - Lawfulness of the processing implies the existence of a legal basis, and a processing that respects the requirements of this legal basis. The control of access to the hate speech database and to the report storage module, and the documentation of these accesses, already required as a safeguard in a previous step of the current PIA, may ensure such a respect, in addition to the planification of a regular independent supervision of the use of these databases. More generally, the demonstration of the compliance of processing activities with the GDPR and the E.U. Directive regulating the

processing of personal data in the police and criminal justice sector - which include the requirements of fairness and of lawfulness of the processing - will be an obligation in 2018.

- Accuracy, reliability, completeness and up-to-dateness of personal data that may be processed by the system are principles that may pose more difficulties in the context of the MANDOLA technical developments, since they imply a processing of personal data that is involuntary. Personal data are not eventually processed in order to identify natural persons and attribute them some other kind of information, but accidentally, because of the nature of the scanned sources. Therefore, it is neither really possible, nor a task attributed to the systems, to verify if data that may be of a personal nature are accurate, reliable, complete or up-to-date. It would even be not desirable that the systems perform such verifications, since it would enable to identify more precisely individuals, whereas the objectives are elsewhere.

  For these reasons, it would be wise to find alternative safeguards, as legislation authorises it[94]. Disclaimers relating to the relative reliability of the dashboard results have already been advised, and the current discussion highlights their importance. In addition, it appears important that disclaimers are also shown to persons who access the hate speech database and the report storage module, highlighting the potential unreliability of hosted data due to the nature of the system, to the nature of information sources, and to the nature of the information itself, in addition to the prohibition to use the database content in order to identify a particular person.

  Moreover, in order to palliate the lack of data's reliability, regular deletion of texts that might contain personal sentences or signs should be ensured, ideally automatically. Such a recommendation has already been made in the previous steps of the current assessment, which shows the importance of time limitation (which is also a requirement in itself as analysed two subsections below).

- **Data minimisation**

Personal data that are potentially processed must be adequate, relevant and not excessive (i.e. limited to the minimum necessary) in relation to the purposes for which they are processed. It seems that, if conditions and recommendations of use are respected, these principles are respected to the utmost possible extent. Indeed, personal data are not collected voluntarily in the MANDOLA hate speech database, which means that are collected only those that are included in the scanned documents and not recognised as potential personal data to be removed (i.e. the name of the user who sent the text and Twitter mentions). These documents are selected because they might correspond to the MANDOLA definition of potentially illegal speech, but not all documents of this kind are kept, whereas the documents that do not correspond to this definition are not collected. In addition, a regular deletion of texts should even more reduce the risk of storing personal data or an important number of such data.

[94] Art. 11, §1 of the GDPR states that "*If the purposes for which a controller processes personal data do not or do no longer require the identification of a data subject by the controller, the controller shall not be obliged to maintain, acquire or process additional information in order to identify the data subject for the sole purpose of complying with this Regulation*". This principle was already implicit in the former legislation. On this issue see the MANDOLA deliverable D2.2 - *Identification and analysis of the legal and ethical framework*, version 2.2.4 of 12 July 2017, MANDOLA project (Monitoring ANd Detecting OnLine hAte speech) - GA n° JUST/2014/RRAC/AG/HATE/6652, http://mandola-project.eu/.

As regards the report storage module, third parties are supposed to only store in it the information provided voluntarily by the users of the smartphone app, reduced (for what regards data that are or might be personal) to an URL, a hate-speech containing text, the device ID if the user agreed to send it, and the content of an optional title for the report. These data are the ones that are required in order to enable analysts to assess the alleged illegal content, and to send their feedback to the author of the report. However, once this feedback has been given, the device ID of the user should be deleted, as well as any personal information included in the title of the report. In addition, after processing and forward to the relevant law enforcement services and eventually other legitimate partners[95], reported contents that might contain personal data and URLs related to stored texts should also be deleted.

- **Time limitation**

In addition to be a safeguard required to protect some other legal requirements we have previously analysed, time limitation is also a legal requirement in itself. This also argues in favour of the study, drafting and implementation of a deletion policy by operators of data sets aiming at training the hate speech classifier, and by third parties receiving reports in the report storage module. This policy should be accompanied with procedural measures ensuring that time limits are observed, and subject to periodic review of the need for the storage of data, in order to ensure it fits with the evolution of the processing's context. Data that are blocked instead of erased should only be processed for the purpose which prevented their erasure, and by a specially authorised person.

- **Appropriate legal ground**

Legal grounds for processing might be different depending on the entity or person who processes personal data.

- ***Data set operators and private database operators***

  The legal ground for (potentially) processing personal data, in relation with processing operations performed by the operators of the data sets used to train the hate speech classifier and by the operators of the hate speech database, is the "*the legitimate interests pursued by the controller (…)*", which may be evoked "*except where such interests are overridden by the interests for fundamental rights and freedoms of the data subject*", on the basis of article 7, f of Directive 95/46/EC and of the national provisions that have implemented this article (the GDPR contains equivalent provisions in its article 6 f).

  This legal ground is also the one of assistance services that are not belonging to the police and justice sector in relation to the texts containing hate speech and their URLs, unless they are on their territory recognised as performing "*a task carried out in the public interest or in the exercise of official authority vested*" in their service (Art. 7, e of Directive 95/64/EC; Art. 6, e of the GDPR). In relation with the data provided by the author of the report, the legal basis will be the consent of this person (Art. 7, a of Directive 95/64/EC; Art. 6, a of the GDPR).

[95] For instance, assistance services part of the INHOPE network use (where law authorises it) to also send the report to the relevant assistant service, if the content is hosted in its territory, and to the hosting provider, where law imposes an action from the latter.

In order to be granted with the benefit of the "legitimate interest pursued by the controller" legal basis, several criteria must be met on the basis of the tests which performance is recommended by the Article 29 Data Protection Working Party[96]:

- ✓ Processing operations must be necessary to pursue the system's purposes. On this field, we have already analysed that collected and processed data appear to be limited to the utmost extent, provided that the recommendations we made so far are followed.
- ✓ As already discussed, the technical systems' purposes are lawful and present a real and present interest[97]. They have moreover been clearly articulated[98].
- ✓ As already discussed, processed data may be sensitive, and certain processing operations may impact some individuals' rights and go beyond data subjects' reasonable expectations, essentially where data from social networks are processed[99]. In addition, data controllers have a dominant position.
- ✓ However, the scale of the processing is very limited, and precautions have been taken in order to ensure that no personal data is used in order to identify an individual or to take a decision against an individual. Given the fact that the MANDOLA technical systems are susceptible to enhance the combat against hate speech, the balance between this interest and the harm suffered by individuals due to these systems - which appears to be very limited in practice - seems to be in favour of the MANDOLA systems (and therefore the controllers') legitimate interests.

- o ***Law enforcement agencies***

Third parties connected to the smartphone app could also be LEAs in charge of processing reports relating to online illegal content. In such case, the legal basis of the processing will be the performance of a task carried out by a competent authority, on the basis of domestic law[100] and, in 2018, on the basis of Article 8 of the Directive on personal data protection for the police and criminal justice sector. Within this context, processing operations must be "necessary" to the performance of this task, which must be understood as referring to the ECHR principles of necessity and proportionality[101].

We have already analysed that the systems at stake may be considered necessary and proportionate, providing that our recommendations linked to these principles are followed, in order to contribute to the combat against online hate speech, which is a

[96] Article 29 Data Protection Working Party, Opinion 06/2014 on the notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC (WP 217), 9 April 2014, III.3.1, p. 25, http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp217_en.pdf (last accessed on 23 May 2014).

[97] See above the sub-section relating to the privacy and personal data protection requirements at the Council of Europe level (necessity and proportionality principles).

[98] See above the sub-section relating to the requirement of legitimate, explicit and specified purpose.

[99] See above the sub-section relating to the requirement of legitimate, explicit and specified purpose.

[100] Based on Article 7 of Directive 95/46/EC in countries where this Directive applies to LEA activities.

[101] See the MANDOLA Deliverable D2.2 - *Identification and analysis of the legal and ethical framework*, version 2.2.4 of 12 July 2017, MANDOLA project (Monitoring ANd Detecting OnLine hAte speech) - GA n° JUST/2014/RRAC/AG/HATE/6652, http://mandola-project.eu/, Section 4.2.3.3.

task carried out by LEAs. We have also already analysed that a legal basis might be necessary to authorise this processing and to provide for adequate safeguards, in case the technical functions of the smartphone app would evolve.

- ***Users of the smartphone app***

  Users who will run the smartphone app on their device will be the ones who will store most of the personal data involved in the system's processing operations. These data processing activities should not be covered by the data protection legislation since they correspond to purely personal activities - as soon as they do not go beyond this perimeter (Art. 3, 2, §2 of Directive 95/46/EC; art. 2, c of the GDPR). This does not mean that developers and broadcasters of the app are exempt from liability as regards the assistance they provide to the user in terms of securing his or her own personal data processing activities (as a minimum through the provision of information in relation to the product they propose, as we will analyse it further in this report[102]).

As a conclusion, provided that our recommendations are applied, potential personal data processing appear to be based on an appropriate legal ground.

- **Data subject information**

Users of the smartphone app must at least be informed of the recipients (which will also be data controllers) of their information, of the purposes of the processing and of the existence of their right of access and communication, in case they decide to send their device ID or other personal information in the report title. This argues in favour of our preceding recommendation to make available, before any report and from the reporting window, the name of the recipients, and to enable the user to choose them. For the rest, developers of the app should find a way to display a link on the detailed personal data policy of each of these recipients (which will have to be clear and consistent and include from 2018 the GDPR or Police Directive requirements - Articles 13), in addition to clear and consistent information relating to the purposes of the processing, to the data subject's right of access, communication and erasure, and to the contact points to be used in this regard. Ideally (using a privacy by design approach), the right of access could be exercised through the app.

In relation to the data that might be related to other persons including hate-speech victims, included in the analysed and stored web or social networks contents, data subject's information is not possible, since the latters are not identified. In this context, necessity and proportionality of the processing operations must be ensured through the implementation of alternative safeguards, such as a clear and visible information relating to these processing operations, their data controller, their purposes, and measures that have been implemented to ensure the confidentiality, the security and the deletion of potentially personal data. This information should ideally be available in all supports of the MANDOLA outcomes, including the MANDOLA website, the MANDOLA reporting portal, information provided through the smartphone app and the MANDOLA dashboard.

[102] See below "Security and confidentiality of the processing".

- **Data subjects' rights of access, communication, rectification and erasure**

Data subject's rights in this regard will have to be ensured by third parties receiving reports.

The same conclusion applies to entities other than LEAs which will store web-contents in the hate speech database and in the report storage module. However in this case, it might be impossible for these entities to answer positively to such a request if the concerned database does not include a functionality that enables to search for a specific word in the stored information, and if findings cannot be corroborated by additional information (such as an URL), in order to avoid communication of information to the wrong person. In addition, it does not seem relevant to advice the implementation of such a searching function, since it would favour persons' identification, whereas the system does not pursue this aim (keeping in mind that entities other than LEAs and the judiciary are not entitled to process personal information relating to penal offences).

In relation with the report storage module and subsequent data collection performed by third parties that would be LEAs after the opening of an investigation, data subjects will have to be granted with rights of access in compliance with domestic law, and, from May 2018, with Articles 14 *et seq.* of the Directive on personal data protection for the police and criminal justice sector.

- **Prohibition of decisions taken on the solely basis of a data processing**

The MANDOLA smartphone app precisely aims at taking decisions that might produce legal effects concerning some persons potentially identifiable in the database, namely perpetrators of hate speech offences. Such decisions (generally of opening of an investigation) will be taken by relevant LEAs, possibly partly on the basis of the MANDOLA reporting system. As a consequence, in order to comply with law, the MANDOLA consortium must make very clear that such decisions cannot be made on the solely basis of this automated processing, and that information must be previously corroborated by other information, external to the system.

Other MANDOLA technical developments do not aim at taking decisions against individuals, only at taking decisions relating to contents and geographical areas, more precisely at classifying some contents as potentially illegal hate speech, and at identifying the level of hate-speech usages in countries and cities. However, such decisions can affect individuals, such as the inhabitants of the related countries or cities. This argues for a particular care in the definition of decision criteria and in the implementation of disclaimers relating to the dashboard's results weaknesses, including their relative reliability, already recommended.

In addition, the MANDOLA non-technical developments aiming at providing information to several categories of recipients, even if they do not contain any personal data, might lead to decisions against individuals or groups or individuals, particularly as they relate to the behaviours to adopt when facing a potentially illegal content, to the appropriate definition of hate speech and to the places where policies against hate speech should be ideally implemented. Even if this issue is going slightly beyond the question of the prohibition of

automated decisions, since no personal data is involved, it is strongly linked to it[103] and must also be a constant concern. In this regard, our recommendations to make recommendations as objective, exhaustive and referenced as possible, with appropriate disclaimers where a given information might encourage behaviours infringing fundamental rights, take on here particular importance.

- **Enhanced protection of some sensitive data**

Sensitive personal data (data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs...), as well as location data (where published in plain text by Internet users on Internet public parts), may be processed by the data set that trains the hate-speech classifier and by third parties as part of the reports they receive. This is due to the nature of the system (which aims at detecting potential penal offences) and to the nature of the data used by this system (web and social media content).

However, such processing of personal data is supposed to be very rare in the dashboard hate speech database since user-names and Twitter mentions are removed from texts that are stored. As regards the report storage module, the smartphone app does not send any Internet content user name, only the text that might contain hate speech (but which might however contain information relating to persons, especially hate-speech victims[104]). In addition, personal data are not especially searched for, and are not used in order to qualify a person whose identification would be desired. In addition, if our recommendations are followed, it will be very clear for all the users of the system that data might be non-reliable, and cannot lead to conclusions in relation with a particular individual.

As a consequence, it appears that all appropriate safeguards have been taken in order to avoid the processing of sensitive data to the utmost possible extent.

For the rest, no location data identified as such, no communications and no traffic data are processed.

- **Security and confidentiality of the processing**

The dashboard's hate speech database (which is associated with the data set that may be used to train the hate speech classifier, and which is a module that is independent from the dashboard as already analysed) is hosted at UCY and benefits from several security measures, including data encryption, password protection and the use of HTTPS. We do not have detailed information relating to other data sets that might alternatively be used. It is however important that they protect the potentially hate speech texts and URLs with appropriate technical and organisational measures in order to ensure a level of security appropriate to the risks[105], including the protection of potential personal data against alteration, unauthorised disclosure or access, and against all other unlawful forms of processing (on the opposite particular protection against accidental or unlawful destruction or accidental loss, mentioned in both the current and the new legislation as a factor of risk,

[103] Indeed, the data protection legislation is a practical implementation of the ECHR principles, applied to personal data processing. The fact that this legislation does not regulate processing operations other than those that include personal data does not mean that the ECHR principles do not apply to the latter processing operations.

[104] This clarification has been added following consultation of the Mandola Advisory Board members.

[105] This wording is used in the new E.U. General Data Protection Regulation. However, this is a traditional requirement in the area of risk management.

does not appear to be required here since ideally all personal data should be removed). Amongst these measures should lie the ones we advised in the above Section relating to legitimate purposes.

Third parties that will receive reports from the smartphone app must implement the same measures, including protection against destruction or loss since partial information might lead to enhance the risk of wrong decisions concerning reports. For the rest, communications between the smartphone app and third parties are encrypted through the https protocol.

In addition, it would be required, in order to prevent any risk in case personal data would be accidentally processed, and in order to protect URLs, that developers of the systems that support the data set and the report storage module implement technical measures that enable the authentication of persons who access the databases, and the retention of logs of access. The personnel of the entities that operate these systems should only be authorised to access these databases on a need-to-know basis, for example in order to perform specific needed tasks such as maintenance or hate speech validation, under confidentiality and purpose non-diversion agreements. Regular control of past accesses should be planned.

Moreover, device IDs should be stored in a separate database to be accessed only for duly justified reasons, with application of the security measures referred to above.

Finally, regarding the content hosted on the user's smartphone, further developments of the app should ensure its protection and the protection of data according to the state of the art. If such a protection was not offered, clear notice should be given to the user in relation to the risks that might be generated by the installation of the software.

- **Data protection authority supervision**

Personal data processing (even potential) performed at the level of the data set and hate speech database on the one hand, and at the level of the report storage modules on the other hand, will have to be notified to relevant data protection authorities in compliance with data controllers' legislations. From May 2018, this obligation will be replaced by an obligation of recording and documentation of processing activities and of compliance with law, in addition to an obligation to notify personal data breaches.

Consultations with relevant supervisory authorities prior processing might also be requested. Indeed this obligation will apply in case the processing would result in a high risk in the absence of measures taken by the controller to mitigate the risk, and it appears that high risks could result - *inter alia* - from an intrusion in the smartphone app hosted on users' devices, and from a diversion or an extension of purposes and technical functions of the dashboard.

- **Liability and accountability of the data controller**

Data controllers in relation with (1) data sets and hate speech databases, and with (2) report storage modules and the processing of reports, will have to ensure compliance with

the data protection legislation, which will be strengthened under the future E.U. legislation[106].

- **Adequate level of protection in some case of data transfers**

The MANDOLA system does not enable, at this stage, any transfer of data to third parties, others than the reports that the user of the smartphone app sends voluntarily. In addition, in the current state of the situation, assistance services against hate speech that are considered to be the future reports recipients are located within the EU. However, as a precaution in case the list of recipients would include services that operate in countries were the level of protection might be non-adequate, it should be recommended to future developers of the smartphone app to warn appropriately the user of this app on the lower state of protection of personal data in certain countries, providing for example a link on the European Commission decisions on the adequacy of the protection of personal data in third countries[107], and advising the user to consult carefully the personal data protection policy of the assistance service that is proposed as recipient of his or her report.

- **Summary of recommendations**

Legal and ethical compliance

- Recommendations that follow are not intended to recall exhaustively legal requirements, but to advise the implementation of safeguards that, in the particular context of the further development or of the use of the MANDOLA products, ensure completion with law where the latter is too general or is difficult to apply comprehensively.
- Measures of awareness raising, control and enforcement must ensure that no modification of the conditions of use of the MANDOLA technical developments (such as they are described in the current report and other MANDOLA technical deliverables) are tolerated, notably at the occasion of the integration of another component to the systems, or/and by using these systems in order to identify individuals or in order to process voluntarily personal data, without performing a new identification of the appropriate safeguards that must be implemented (the best method being the performance of a PIA that would take into account these modification and would follow again each step of the PIA method proposed in the MANDOLA Deliverable D2.4a), since the conclusions of the current PIA as a whole and of each of its steps would not be adapted anymore to such a new situation.

Data protection authorities' supervision

- Consultations with relevant supervisory authorities prior processing might be requested from 2018. Indeed this obligation will apply in case the processing would result in a high risk in the absence of measures taken by the controller to mitigate the risk, and it appears that high risks could result - *inter alia* - from an intrusion in the

[106] See the MANDOLA deliverable D2.2 - *Identification and analysis of the legal and ethical framework*, version 2.2.4 of 12 July 2017, MANDOLA project (Monitoring ANd Detecting OnLine hAte speech) - GA n° JUST/2014/RRAC/AG/HATE/6652, http://mandola-project.eu/, Section 4.2.3.3.13.

[107] http://ec.europa.eu/justice/data-protection/international-transfers/adequacy/index_en.htm.

smartphone app hosted on users' devices, and from a diversion or an extension of the purposes and technical functions of the dashboard.

Information of Internet users

- A clear, consistent and visible information should be provided in relation to the purposes of the monitoring dashboard and of the smartphone app, to the data controllers and contact details of data protection officers where applicable, to the processing operations and purposes which are authorised to third parties that will receive reports through the app, to the exact nature of potentially personal data that may be included in databases, to data sources, to their right of access, communication and erasure and the contact points to be used in this regard, and to measures put in place in order to ensure the protection of privacy and personal data (including confidentiality, security and deletion).

  Ideally, this information should be available in all supports of the MANDOLA outcomes, including the MANDOLA website, the MANDOLA reporting portal, information provided through the smartphone app and the MANDOLA dashboard. It should also be included in all the supports and channels that will give access to the MANDOLA dashboard results.

- In particular, users of the smartphone app must at least be informed of the recipients (which will also be data controllers) of their information, of the purposes of the processing and of the existence of their right of access and communication, in case they decide to send their device ID or other personal information in the report title, and on the contact point to be used in this regard.

  To this end the future developers of the app should in particular make the necessary as to enable the users to visualise the name of the recipients of their report, before transmission of the latter, and to choose to remove some recipients and/or to add ones. They should find a way to display a link on the detailed personal data policy of each of these recipients (which will have to be clear and consistent and include from 2018 the GDPR or Police Directive requirements - Articles 13), in addition to clear and consistent information relating to the purposes of the processing, to the data subject's right of access, communication and erasure, and to the contact points to be used in this regard. Ideally (using a privacy by design approach), the right of access could be exercised through the app.

Anonymisation

- Ideally, no individual should be identifiable in the MANDOLA hate database and, after transmission to relevant LEAs, in the report storage modules. This implies to remove all names and other visible signs that might lead to or that might be personal data.

- The mechanism that removes a part of the geolocation coordinates in order to anonymise data collected and shown in the monitoring dashboard is of utmost importance and must be particularly preserved and secured against removal or circumvention.

- Once the smartphone owner has received feedback from the recipient of his or her report, his or her device ID should be deleted, as well as any personal information included in the title of the report.

- After the processing of reports and their transmission to the relevant law enforcement services and eventually other legitimate partners[108], reported contents that might contain personal data and URLs related to stored texts should be deleted by the controllers of the report storage modules.

Security

- Since simple texts might occasionally lead to identify an individual, even if all (technically) visible personal data are removed, as well as URLs each time they are kept in relation to a given content, appropriate technical and organisational measures against undue internal or external access must be applied in order to ensure that (1) access to a text or to a URL stored in the hate speech database pursues the solely aim of verifying the illegal nature of a given content, in order to enhance the performances of the dashboard, and to ensure that (2) access to a text or an URL stored in a report storage module pursues the solely aim of assessing the illegal nature of the content, in compliance with domestic law and the policies of the assistance service, and of forwarding it to competent authorities.
    - To this end technical (to be implemented by future developers of the hate speech database and of the report storage module) and organisational (to be implemented by data controllers) measures should ensure that access to contents and to URLs is restricted to identified persons accredited to do it on a "need to know" or "need to use" basis in order to perform specific needed tasks (such as maintenance or hate speech validation), under agreements of confidentiality and purpose non-diversion. Access control and record of access should be in place as well as a regular independent supervision of past accesses and of their purposes.
    - Recourse to hosting providers should be avoided, and where impossible strong contractual and security measures should prevent any undue access, modification, record or other processing of data by a hosting provider or a technical provider which services would be used by the operators of the hate speech database or of the report storage module.
    - Device IDs should be stored in a separate database to be accessed only for duly justified reasons, with application of the security measures referred to above.
- Further developments of the smartphone app should ensure the protection of the content hosted on smartphones. If such a protection was not offered, clear notice should be given to the users in relation to the risks that might be generated by the installation of the software.

Prevention of discrimination and of arbitrary decisions

- Persons who access the hate speech database and the report storage module should also see disclaimers highlighting the potential unreliability of hosted data due to the nature of the system, to the nature of information sources, and to the nature of the

[108] For instance, assistance services part of the INHOPE network use (where law authorises it) to also send the report to the relevant assistant service, if the content is hosted in its territory, and to the hosting provider, where law imposes an action from the latter.

information itself, in addition to the prohibition to use the database content in order to identify a particular person.

- The fact that a name of person or a sign/a sentence that might identify a person, included in the database, will never correspond for sure to a real and identifiable natural person, must be very clear for any user of the MANDOLA hate speech database and of the report storage module.
- As already recommended, a disclaimer (visible where the results per country or per city and the results of the Hate strength gauge are displayed) should detail very clearly the variables that are taken into account in order to calculate the hate speech score of countries and cities (such as the number of inhabitants and the volume of Internet content produced each day), avoid the use of the word "dangerous"[109] and explain on the opposite in simple terms that these statistics cannot represent the state of dangerousness of a given country or city, in particular since (1) they don't take into account several important factors such as the Internet penetration, the number of Internet users and the frequency and their habits in terms of Internet usage; (2) the MANDOLA dashboard shows hate speech that is potentially illegal in one or several E.U. countries but that might not be illegal in one or several others; (3) the context of the speeches are not taken into account and the assessment of contents is not exact science; and (4) even a high level of illegal online hate speeches (which might be produced by the same group of persons, and which are eased by the simplicity of posting on the Internet) does not necessarily means that a given country or city as a whole is dangerous in terms of hate speech usage.

  In particular, strategic decisions taken on the basis of the MANDOLA monitoring dashboard should not restrict some individual's rights. If such a restriction of rights is planned, these decisions should not be taken if not corroborated by additional information obtained from outside the MANDOLA system.
- In the same line, the MANDOLA recommendation of use must make very clear that decisions that might produce legal effects concerning persons potentially identifiable in the database, perpetrators of hate speech offences at the first place, cannot be made on the solely basis of this automated processing, and that information must be previously corroborated by other information, external to the system.
- As already recommended, the MANDOLA outcomes taking the form of information provided to policy makers, to the industry and to Internet users must be as objective, exhaustive and referenced as possible, with appropriate disclaimers where a given information might encourage behaviours infringing fundamental rights.

Time limitation

- Deletion policies should be implemented by operators of data sets aiming at training the hate speech classifier, and by third parties receiving reports in the report storage module. These policies should organise the regular deletion of URLs linked to texts and of all the texts that might contain indirect personal data, as long as they are not absolutely useful to the proper functioning of the system. These policies should be accompanied with procedural measures ensuring that time limits are observed, and

[109] This specification results from the consultation of the Mandola Advisory Board members.

subject to periodic review of the need for the storage of data, in order to ensure it fits with the evolution of the processing's context. Data that are blocked instead of erased should only be processed for the purpose which prevented their erasure, by a specially authorised person.

Data quality and data subjects' rights of access, communication, rectification and erasure

- It does not seem necessary to recommend that a function is created in order to enable the search, in the hate speech database and in report storage modules, for a name of a person in order ensure data quality and to enable data subjects to exercise their rights of access, communication, rectification and erasure. Indeed, it would favour persons' identification, whereas the system does not pursue this aim (keeping in mind that entities other than LEAs and the judiciary are not entitled to process personal information relating to penal offences).

Protection against data transfer in countries that do not ensure adequate level of protection

- As a precaution in case the list of recipients of the smartphone app would include services that operate in countries were the level of protection might be non-adequate, it should be recommended to future developers of the smartphone app to warn appropriately the user of this app on the lower state of protection of personal data in certain countries, providing for example a link on the European Commission decisions on the adequacy of the protection of personal data in third countries[110], and advising the user to consult carefully the personal data protection policy of the assistance service that is proposed as recipient of his or her report.

3.3.1.6 Identification of the threat sources

Threat sources will not be all identifiable for sure in the current PIA, since they need to be identified taking into account the concrete context in which the MANDOLA systems will be implemented.

Therefore, the following table is an attempt to identify most of the threat sources that may threaten personal data or privacy in most contexts. Future data and systems controllers might have to refine this analysis (adding or removing some threat sources and / or threat examples) taking into account the specificities of their own structure and context.

Identification of threat sources		
Types of threat sources	**selected or not**	**Example *(example of action)***
Human source, internal, malicious, with low capacities (network analysis, hate speech database, monitoring dashboard, reports processing and storage).	Yes, some data or system controllers might be exposed to this threat.	Maintenance / cleaning staff *(accessing paper copies of reports --> publication; or damaging the computer system)*. Analyst with low level of access authorisation / Trainee acting playfully or in order to strengthen the combat against online hate speech, while breaching data protection

[110] http://ec.europa.eu/justice/data-protection/international-transfers/adequacy/index_en.htm.

		rules *(search for the direct or indirect identity of the author of a potentially illegal Tweet --> paper copy or publication).*
Human source, internal, malicious, with low capacities (app on smartphones and communication of report).	Yes, smartphone users might be exposed to this threat.	Person belonging to the user's environment, wishing to prejudice this user *(persuading him/her to not use the app/to not publish on social networks).*
Human source, internal, malicious, with low capacities (Internet users' freedoms).	Yes, Internet users might be exposed to this threat.	Person belonging to the user's environment, wishing to prejudice this user *(persuading him/her to not publish on social networks; reporting him/her as author of hate speech content).*
Human source, internal, malicious, with significant capacities (network analysis, hate speech database, monitoring dashboard, reports processing and storage).	Yes, some data or system controllers might be exposed to this threat.	Analyst initiative aiming to strengthen the combat against online hate speech, while breaching data protection rules *(removal of functions protecting users' names and geolocations)* or willing to prejudice MANDOLA outcomes *(update with false information of the MANDOLA publications)*. Hotline analyst with technical knowledge wanting to harm the organisation or the combat against hate *(injection of false data into the system).* Sub-contractor, provider, help-desk agent *(no respect of the contract that imposes no data access and/or reuse - personal inquiry / publication of contents and URLs).*
Human source, internal, malicious, with significant capacities (app on smartphones and communication of report).	Yes, smartphones users might be exposed to this threat.	Person belonging to the user's close environment, knowing the smartphone access code or having the possibility to consult the smartphone, willing to prejudice this user or monitor his/her *activity (sending of false reports; access to previous reports in a familiar cell where retaliation is possible - including if previous reports were relating to this family activity).*
Human source, internal, malicious, with significant capacities (Internet users' freedoms).	Yes, Internet users might be exposed to this threat.	Person belonging to the user's close environment, with significant technical skills willing to prejudice this user *(publishing false hate speeches under the user's name).*
Human source, internal, malicious, with unlimited capacities (network analysis, hate speech database, monitoring dashboard, reports processing and storage).	Yes, some data or system controllers might be exposed to this threat.	Head of the organisation whishing strengthening the combat against online hate speech, while disregarding data protection rules *(removal of functions protecting users' names and geolocations)*. System administrator acting in a spirit of revenge or pursuing personal interests *(injection of false data into the system).* Developer *(creating a weakness in the system in order to later on compromise it*).
Human source, internal, malicious, with unlimited capacities (app on	Yes, smartphones users might be	Person belonging to the user's close environment, knowing the smartphone

smartphones and communication of report).	exposed to this threat.	access code or having the possibility to consult the smartphone, having high technical skills and willing to prejudice this user or monitor his/her activity *(deletion of personal information preventing the exercise of the right of access).*
Human source, internal, malicious, with unlimited capacities (Internet users' freedoms).	No, Internet users are not likely to be exposed to this threat.	
Human source, external, malicious, with low capacities (network analysis, hate speech database, monitoring dashboard, reports processing and storage).	Yes, some data or system controllers might be exposed to this threat.	Person(s) without important technical skills *(sending several times the same report in order to bias results).*
Human source, external, malicious, with low capacities (app on smartphones and communication of report).	Yes, smartphones users might be exposed to this threat.	Vandal with very common technical skills *(smartphone theft--> password too simple, found --> access to reports).*
Human source, external, malicious, with low capacities (Internet users' freedoms).	Yes, Internet users might be exposed to this threat.	Other user of the Internet service wishing to prejudice someone *(false reports to the MANDOLA system, which can be perceived as hate speech if taken out of their context (ex. theatre scenario).*
Human source, external, malicious, with significant capacities (network analysis, hate speech database, monitoring dashboard, reports processing and storage).	Yes, some data or system controllers might be exposed to this threat.	Activist or NGO guided by ideology / politic beliefs; passionate hacker; former agent whishing harming the organisation or pursuing personal interests *(compromising the system or having kept devices ID; modification of MANDOLA publications - providing wrong advices as a result).*
Human source, external, malicious, with significant capacities (app on smartphones and communication of report).	Yes, smartphones users might be exposed to this threat.	Vandal with important technical skills *smartphone theft--> password found --> access to reports --> anonymous publication under the user identity).* Activist or NGO guided by ideology / politic beliefs; passionate hacker; former agent of a reporting platform whishing harming the organisation or the smartphones users or pursuing personal *interests (knowing how compromising smartphones through the app, or smartphones of which the ID is known).*
Human source, external, malicious, with significant capacities (Internet users' freedoms).	Yes, Internet users might be exposed to this threat.	Hateful persons or group, including terrorists *(manufacturing of false hate speech content + anonymous reports -> investigations against innocent people).*
Human source, external, malicious, with unlimited capacities (network analysis, hate speech database, monitoring dashboard, reports processing and storage).	No, data or system controllers should unlikely be exposed to this theat.	Criminal organisation *(acting on the premises).*
Human source, external, malicious, with unlimited capacities (app on smartphones and communication of	Yes, smartphones users might be exposed to this	Staff of an assistance service wishing to prejudice authors of reports *(action on smartphones using the app; collection and re-*

report).	threat.	*use of their device IDs).*
Human source, external, malicious, with unlimited capacities (Internet users' freedoms).	Yes, Internet users might be exposed to this threat.	Internet service provider's personnel wishing to prejudice the user or willing to remove chocking but non-illegal content *(closure of the user's account based on the publication of a chocking - alleged illegal - content).*
Human source, internal, without malicious intent, with low capacities (network analysis, hate speech database, monitoring dashboard, reports processing and storage).	Yes, data or system controllers might be exposed to this threat.	Analyst / trainee with limited awareness or limited motivation or acting unconsciously *(assessing contents wrongfully)*.
Human source, internal, without malicious intent, with low capacities (app on smartphones and communication of report).	Yes, smartphones users might be exposed to this threat.	Person belonging to the user's environment, proving bad advice *(persuade him or her to not use the app due to the privacy limitation it incurs).*
Human source, internal, without malicious intent, with low capacities (Internet users' freedoms).	Yes, Internet users might be exposed to this threat.	Person belonging to the user's environment, proving bad advice *(persuade him or her to not publish on social networks due to close monitoring).*
Human source, internal, without malicious intent, with significant capacities (network analysis, hate speech database, monitoring dashboard, reports processing and storage).	Yes, data or system controllers might be exposed to this threat.	Analyst / IT developer with limited awareness or limited motivation or acting unconsciously; agent of a sub-contractor / hosting provider / services provider with limited awareness or limited motivation or acting unconsciously *(collection or non-protection of personal data; display of detailed geolocations and URLs)*. Personnel willing to strengthen the combat against hate *(providing wrong advices through the modification of the MANDOLA publications).*
Human source, internal, without malicious intent, with significant capacities (app on smartphones and communication of report).	Yes, smartphones users might be exposed to this threat.	Person belonging to the user's environment, having access to the smartphone and wishing to protect the user *(removing the app and by mistake the data).*
Human source, internal, without malicious intent, with significant capacities (Internet users' freedoms).	Yes, Internet users might be exposed to this threat.	Person belonging to the user's environment, having access to the smartphone and wishing to protect the user *(asking a provider to remove a text for privacy reasons and as a result obtaining the closure of the account).*
Human source, internal, without malicious intent, with unlimited capacities (network analysis, hate speech database, monitoring dashboard, reports processing and storage).	Yes, data or system controllers might be exposed to this threat.	Head of operations with limited awareness regarding security or personal data protection within the framework of the fight against hate speech *(collection or non-protection of personal data; display of detailed geolocations and URLs).*
Human source, internal, without malicious intent, with unlimited capacities (app on smartphones and communication of report).	Yes, smartphones users might be exposed to this threat.	User him or herself *(self-censorship, mistake in installation/software removal).*
Human source, internal, without malicious intent, with unlimited	Yes, Internet users might be exposed	User him or herself *(self-censorship; wrong attitudes due to a bad understanding or an undue modification of the MANDOLA*

capacities (Internet users' freedoms).	to this threat.	*advices).*
Human source, external, without malicious intent, with low capacities (network analysis, hate speech database, monitoring dashboard, reports processing and storage).	Yes, data or system controllers might be exposed to this threat.	Users with poor awareness on legal issues, thinking to have the duty of finding and reporting hate speech *(abnormally important number of reports relating to legal contents)*.
Human source, external, without malicious intent, with low capacities (app on smartphones and communication of report).	No, smartphone users are unlikely to be exposed to this threat.	
Human source, external, without malicious intent, with low capacities (Internet users' freedoms).	No, Internet users are unlikely to be exposed to this threat.	
Human source, external, without malicious intent, with significant capacities (network analysis, hate speech database, monitoring dashboard, reports processing and storage).	Yes, data or system controllers might be exposed to this threat.	Policy makers *(misunderstanding of statistics - policies attempting fundamental freedoms).* LEA *(opening of an investigation based on wrong information).* Negligence of the personnel of one of the data or system controllers *(non-taking into account of legislative modifications relating to illegal hate speech).*
Human source, external, without malicious intent, with significant capacities (app on smartphones and communication of report).	Yes, smartphone users might be exposed to this threat.	Personnel of one of the data or system controllers *(by negligence: non-deletion of the device ID / other data; maintenance issues leading to obsolescence of the list of recipients or their data protection policies or the dashboard statistics; thinking knowing the truth: modifying MANDOLA information, thereby providing false information / wrong advice).*
Human source, external, without malicious intent, with significant capacities (Internet users' freedoms).	Yes, Internet users might be exposed to this threat.	LE agent (investigating a wrong IP address following a report). Personnel of one of the data or system controllers, by negligence or thinking knowing the truth *(maintenance issues leading to obsolescence or falsehood of the MANDOLA recommendations or dashboard statistics --> Internet users misled on behaviours to be adopted; policy makers misled on decisions to make).* Policy makers *(misled on decisions to make).*
Human source, external, without malicious intent, with unlimited capacities (network analysis, hate speech database, monitoring dashboard, reports processing and storage).	Yes, data or system controllers might be exposed to this threat.	Court decision *(considering a specific legal basis is required)*.
Human source, external, without malicious intent, with unlimited capacities (app on smartphones and communication of report).	Yes, Internet users might be exposed to this threat.	Future developers of the app *(app update - undue removal of personal content).*

Human source, external, without malicious intent, with unlimited capacities (Internet users' freedoms).	Yes, Internet users might be exposed to this threat.	Internet service provider *(action on the user's account).* Third party copying the MANDOLA information but modifying the content of advices, thinking knowing the truth *(Internet users misled on behaviours to be adopted; policy makers misled on decisions to be made).*
Malicious code of unknown origin (network analysis, hate speech database, monitoring dashboard, reports processing and storage).	Yes, data or system controllers might be exposed to this threat.	Malicious code untargeted of unknown origin, computer virus *(reaching servers / data controllers' systems).*
Malicious code of unknown origin (app on smartphones and communication of report).	Yes, smartphones users might be exposed to this threat.	Malicious code untargeted of unknown origin, computer virus *(acting on the app/through the app).*
Malicious code of unknown origin (Internet users' freedoms).	Yes, Internet users might be exposed to this threat.	Malicious code untargeted of unknown origin, computer virus *(coming from web supports of the MANDOLA information).*
Natural phenomenon (network analysis, hate speech database, monitoring dashboard, reports processing and storage).	Yes, data or system controllers might be exposed to this threat.	Lightning, other natural phenomenon, wear *(damaging hardware / softwares; obsolescence of recommendations).*
Natural phenomenon (app on smartphones and communication of report).	Yes, smartphones users might be exposed to this threat.	Lightning, other natural phenomenon, wear *(obsolescence of the app, of recommendations).*
Natural phenomenon (Internet users' freedoms).	Yes, Internet users might be exposed to this threat.	Lightning, other natural phenomenon, wear *(obsolescence of recommendations to them / to policy makers).*
Natural or sanitary disaster (network analysis, hate speech database, monitoring dashboard, reports processing and storage).	Yes, data or system controllers might be exposed to this threat.	Sickness *(no control of access to personal information; no manual data deletion in order to respect time limits; maintenance issues leading to obsolescence or falsehood of the list of reports' recipients or of the MANDOLA recommendations or dashboard statistics while a regular update was announced --> Internet users misled on behaviours to be adopted; policy makers misled on decisions to make).*
Natural or sanitary disaster (app on smartphones and communication of report).	Yes, smartphones users might be exposed to this threat.	Sickness of the personal of one of the data or system controllers *(no control of access to personal information; no manual data deletion in order to respect time limits; maintenance issues leading to obsolescence or falsehood of the list of reports' recipients or of the MANDOLA recommendations or dashboard statistics while a regular update was announced --> Internet users misled on behaviours to be adopted; policy makers misled on decisions to make).*
Natural or sanitary disaster (Internet	Yes, Internet users might be exposed	Sickness of the personal of one of the data or system controllers *(maintenance issues*

users' freedoms).	to this threat.	*leading to obsolescence or falsehood of the MANDOLA recommendations or dashboard statistics while a regular update was announced --> Internet users misled on behaviours to be adopted; policy makers misled on decisions to make).*
Animal activity (network analysis, hate speech database, monitoring dashboard, reports processing and storage).	No, data or system controllers are unlikely to be exposed to this threat.	
Animal activity (app on smartphones and communication of report).	No, smartphone users are unlikely to be exposed to this threat.	
Animal activity (Internet users' freedoms).	No, Internet users are unlikely to be exposed to this threat.	
Internal event (network analysis, hate speech database, monitoring dashboard, reports processing and storage).	Yes, data or system controllers might be exposed to this threat.	Fire, computer failure *(damaging the information system)*; change in network infrastructure, addition of a new component to the information system *(disrupting the information system - causing maintenance issues due to poor awareness of new staff- see natural disaster for example of possible consequences).*
Internal event (app on smartphones and communication of report).	Yes, smartphones users might be exposed to this threat.	Defect of the app *(information deletion).*
Internal event (Internet users' freedoms).	No, Internet users are unlikely to be exposed to this threat.	

Table 1: Identification of threat sources

3.3.2 Identification of the assets

The aim of this step is to identify the assets that are included in the scope of the study. Assets are the goods, resources or values, including immaterial, that need to be protected, and to the elements that support them, which might especially be human, an hardware component or a software.

3.3.2.1 Primary assets

Primary assets, in other words the non-material resources that need to be protected (and therefore whose availability, integrity and confidentiality must be ensured) are identified in the current PIA as being the following:

- Personal data, potentially personal data and URLs, processed by the system or that will result from the system processing's operations (data might be relating to users of the smartphone app, to authors of hate speech reports, to victims and potential

perpetrators of hate speech, and to possibly any other person referred to in an Internet text);

- Compliance with legal requirements that must be respected based on the ECHR and on the Data Protection Legislation, at the exception of the requirement of security and confidentiality[111], namely[112]:

  LR1 - Specific, clear, accessible, stable and foreseeable legal basis.

  LR2 - Necessity of the project or processing (efficient answer to a pressing social need).

  LR3 - Proportionality of the answer (strict necessity taking into account the severity of the social need, the proportionality of the restricted behaviour, the scope of the interference (especially in terms of number of data, of people and places affected, of situations of use of the system/project), and the nature of other answers available).

  LR4 - Legitimate, specified, explicit, compatible and non-diverted purpose.

  LR5 - Data quality (fair and lawful processing of accurate, reliable[113] and up-to-date data).

  LR6 - Minimisation (adequate, relevant and limited to what is necessary[114] to reach the purposes).

  LR7 - Time limitation (kept for no longer than necessary to reach the purposes), including by the setting-up of time limits "*for erasure or for a periodic review*"[115].

  LR8 - Data subject consent or other legal ground listed by law.

  LR9 - Data subject information.

  LR10 - Data subject rights of access, communication, rectification and erasure.

  LR11 - Prohibition of decisions based solely on automated processing where they produce legal effects concerning the data subject or similarly significantly affect him or her[116].

  LR12 - DPA notification where required by law.

  LR13 - Controller's accountability (measures ensuring and demonstrating legal compliance including privacy by design and default; record of processed activities; DPIA where needed)[117].

[111] The legal requirement of security and confidentiality of the processing is not included in the following assets, since security and confidentiality measures aim at ensuring both the efficiency of most of other legal requirements and the security of the system, including the processed data. As a consequence, security and confidentiality measures will be presented in steps 3.3.2.4 and 3.5 of the current PIA.

[112] In relation to the content of these requirements, see Estelle De Marco *et al.*, MANDOLA deliverable D2.2 - *Identification and analysis of the legal and ethical framework*, MANDOLA project (Monitoring ANd Detecting OnLine hAte speech) - GA n° JUST/2014/RRAC/AG/HATE/6652, http://mandola-project.eu/, 12 July 2017.

[113] Keeping in mind that the non-reliability of data, which might occur where collecting data on the Internet, must lead to implement appropriate safeguards aiming at protecting individuals' rights.

[114] This formulation is the one of the GDPR. Directive 95/46/EC evokes "non excessive" data, which in essence has the same meaning.

[115] GDPR, recital n°39.

[116] Article 22 of the GDPR.

[117] The content of LR13 is anticipating the introduction of the GDPR.

LR14 - Enhanced protection of sensitive data (special categories of data listed in the law, in addition to communications, location data and traffic data).

LR15 - Adequate level of protection in case of certain data transfer outside the EU, in compliance with law.

- Fundamental rights exercised in the private sphere or on the basis of the use, the availability or of the confidentiality of personal data (at the exclusion of those processed by the system[118]), namely[119]:

FR1 - The right to privacy.

FR2 - The right to personal data protection.

FR3 - Freedom of expression.

FR4 - The right to presumption of innocence.

FR5 - The right to non-discrimination.

FR6 - The right to freedom of assembly.

FR7 - The right to freedom of movement.

FR8 - The right to liberty and security.

FR9 - The freedom to conduct a business.

3.3.2.2 Supporting assets

Supporting assets, in other words components (of a technical nature or not) of the information system or more widely of the Society, which support primary assets, and which vulnerabilities may be exploited to harm a primary asset, cannot be identified exhaustively by the MANDOLA consortium since they might be specific to data or system controllers, smartphone users or Internet users. However, a first basic list of these assets may be the following.

- **SYS DEP- Computer and telephone systems (dependent from the system controller)**
  - HAR - Hardware
    - computers used for visualisation (internal),
    - Storing server (internal or externally hosted)
  - SOF - Softwares
    - MANDOLA system components
    - Other softwares used by the organisation for the needs of the MANDOLA processing chain
    - Other softwares eventually used by the organisation, inter alia to transmit internal communications
  - DATA - Information provided by the MANDOLA partners
    - Information to Internet users (dashboard results, reports)

[118] Such exclusion prevents redundancies since impacts on freedoms due to a risk posed to processed personal data will already be studied through the study of the risks posed by the first category of assets (personal data processed or created by the system).

[119] See the MANDOLA deliverable D2.2 - *Identification and analysis of the legal and ethical framework*, version 2.2.4 of 12 July 2017, MANDOLA project (Monitoring ANd Detecting OnLine hAte speech) - GA n° JUST/2014/RRAC/AG/HATE/6652, http://mandola-project.eu/publications.

    - Information to Industry (dashboard results, reports)
    - Information to policy makers (dashboard results, reports)
  - CTC - Computer and telephone channels (cable, wifi, 4G...)

- **SYS INDEP- Computer and telephone systems (Computer and information systems, independent from the system or data controller)**
  - HAR - Hardware
    - Social networks and web servers
    - Smartphones used to welcome the MANDOLA app
    - Devices used by Internet users to access the Internet
  - SOF - Softwares
    - Web and social networks analysed by the project's features
    - Softwares used on smartphones welcoming the MANDOLA app
    - Softwares used on their computer or device by Internet users
  - DATA - Information provided by the MANDOLA partners
    - Information to Internet users (dashboard results, reports)
    - Information to Industry (dashboard results, reports)
    - Information to policy makers (dashboard results, reports)
  - CTC - Computer and telephone channels (cable, wifi, 4G...)

- **ORG - organisation**
  - PER - Persons
    - Analyst/staff members authorised to access the system
    - Providers authorised to access the system
    - Other persons eventually authorised to access the system
  - PAP - Paper copies
    - reproducing MANDOLA results
    - reproducing communications with other parties to the MANDOLA system
    - raising awareness about the conditions of use of the system
    - internal communications
  - ELEC - Electronic copies
    - reproducing MANDOLA results
    - reproducing communications with other parties to the MANDOLA system
    - raising awareness about the conditions of use of the system
    - internal communications
  - CHA - Interpersonal channels
    - Organisational channels and processes
    - Verbal communications

- **ORGPER IND - Organisations and persons, independent from the system or data controller**
  - PER - Organisations and persons
    - Personnel of the controllers of other systems or data processing that are part of the MANDOLA solution
    - Users of smartphones that welcome the app
    - Internet users (fundamental rights' holders)

- Internet service providers and their personnel

- **AUT - National, European and International public authorities**
  - PAR - Parliament
  - JUD - Judiciary
  - IND - Independent authorities
  - ADM - Administrative authorities

Table 2: List of supporting assets

3.3.2.3 Links between primary assets and supporting assets

The following table presents (the basic list of) supporting assets and their (indisputable or potential) links with primary assets

Primary assets / Supporting assets	(Personal) data / URLs	Privacy & DP requirements	Fundamental rights
SYS DEP- Computer and telephone systems (dependent from the system or data controller)	**x**	**x**	**x**
HAR - computers used for visualisation (internal)	x	x	
HAR - Storing server (internal or externally hosted)	x	x	
SOF - MANDOLA system components	x	x	
SOF - Other softwares used by the organisation for the needs of the MANDOLA processing chain	x	x	
SOF - Other softwares eventually used by the organisation, inter alia to transmit internal communications	x	x	
DATA - Information to Internet users (dashboard results, reports)			x
DATA - Information to Internet stakeholders (dashboard results, reports)			x
DATA - Information to policy makers (dashboard results, reports)			x
CTC - Computer and telephone channels (cable, wifi, 4G...)	x	x	
SYS INDEP - Computer and telephone systems (Computer and information systems, independent from the system or data controller)	**x**	**x**	**x**
HAR - Social networks and web servers	x		
HAR - Smartphones used to welcome the MANDOLA app	x	x	x
HAR - Devices used by Internet users to access the Internet			x
SOF - Web and social networks analysed by the project's features	x		x
SOF - Softwares used on smartphones welcoming the			x

MANDOLA app			
SOF - Softwares used on their computer or device by Internet users			x
DATA - Information to Internet users (dashboard results, reports)			x
DATA - Information to Internet stakeholders (dashboard results, reports)			x
DATA - Information to policy makers (dashboard results, reports)			x
CTC - Computer and telephone channels (cable, wifi, 4G...)	x		x
ORG - organisation	**x**	**x**	**x**
PER - -Analyst/staff members authorised to access the system	x	x	x
PER - Providers authorised to access the system	x	x	x
PER - Other persons eventually authorised to access the system	x	x	x
PAP - reproducing MANDOLA results	x	x	x
PAP - -reproducing communications with other parties to the MANDOLA system	x	x	x
PAP - -raising awareness about the conditions of use of the system		x	x
PAP - Internal communications	x	x	x
CHA - Organisational channels and processes	x	x	x
CHA - Verbal communications	x	x	x
OPE INDEP - Organisations and persons, independent from the system or data controller	**x**	**x**	**x**
PER - Personnel of the controllers of other systems or data processing that are part of the MANDOLA solution	x *(hotlines sending reports to train the classifier only)*	x *(idem col. A - minimisation)*	x
PER - Users of smartphones that welcome the app	x	x	x
PER - Internet users (fundamental rights' holders)	x	x	x
PER - Internet service providers and their personnel	x		x
AUT - National, European and International public authorities		**x**	**x**
PAR - Parliament			x
JUD - Judiciary		x	x
IND - Independent authorities		x	x
ADM - Administrative authorities		x	x

Table 3: Links between supporting assets and primary assets

3.3.2.4 Existing security and compliance measures

The prototype delivered by the MANDOLA consortium includes mainly the following security and privacy measures (the extensive list of these measures resulting from the PIA and implemented during research is available in Sections 4.1.1, 4.2.1 and 4.2.2 of the current report):

- Hate data base security (knowing that the hate speech data base is a module that is independent from the dashboard, and that it might only provide the service consisting in training the hate speech classifier, on demand): the database is hosted by UCY on their premises in Cyprus. The hosting server is secured with several security measures including password protection, data encryption and use of the secure communication protocol HTTPS.
- Https communications between the smartphone user and the recipients of reports.
- Security of the reporting portal[120]: the portal is hosted by FORTH on their premises in Heraklion. The hosting server features two Intel Xeon dual-core CPUs running at 2.66GHz and a total memory of 4GB. It is connected to the Internet through FORTH's 10 Gigabit connection to the GRNET backbone. The server has two high performance SAS disks (10k RPM) arranged as RAID-1 for fault tolerance. The server is protected by firewalls and is internally and externally monitored in order to minimize the risk from cyber-threats. Additionally, remote backups through the rsync utility are performed on a daily basis. It is also important that the server resides in a protected physical environment. It is located in one of FORTH's data-centres. For ensuring optimal operating environment, it is equipped with industrial-strength air conditioning with more than 240.000BTUs efficiency. In power emergencies, it is supported by a UPS power supply and an external power generator which is engaged automatically on power failure. Additionally, the data-centre features an automatic carbon dioxide fire-extinguishing system.
- Possibility for the users of the smartphone app to not send their device ID to the recipients of their reports;

3.3.3 Preparation of metrics

Parameters and scales that are used within the framework of the current PIA are the following.

3.3.3.1 Definition of the safety criteria and of the scale of needs

Safety criteria are at least the availability of primary assets, their integrity, and their confidentiality.

As a reminder, primary assets are in the current PIA

- Personal data, potentially personal data and URLs, processed by the system or that will result from the system processing's operations;

[120] See the MANDOLA Deliverable D3.2 - *Reporting Portal*, October 2016, MANDOLA project (Monitoring ANd Detecting OnLine hAte speech) - GA n° JUST/2014/RRAC/AG/HATE/6652, http://mandola-project.eu/, pp. 15-16.

- Legal requirements that must be respected based on the ECHR and on the Data Protection Legislation, at the exception of the requirement of security and confidentiality;
- Fundamental rights exercised in the private sphere or on the basis of the use, the availability or of the confidentiality of personal data.

In order to express safety needs, selected safety criteria are the following:

Safety criteria	Definition
Availability	Availability of primary assets at the desired time for authorised personnel
Integrity	Accuracy and completeness of primary assets
Confidentiality	Accessibility of primary assets to authorised persons only

Table 4: Selected safety criteria

In order to express security needs in terms of availability, the selected scale is the following:

Levels of the scale	Description of the scale
4. 0 h	The primary asset must not be unavailable
3. Less than 24 h	The primary asset must be available within 24 hours
2. Less than 72h	The primary asset must be available within 72 hours
1. Less than 2 weeks	The primary asset can be unavailable more than 72 hours but must be recovered within a reasonable time (maximum of 2 weeks).
0. Unavailable	The primary asset can be definitely unavailable.

Table 5: Selected availability scale

In order to express safety needs in terms of integrity, the selected scale is the following:

Levels of the scale	Description of the scale
4.Integrity	The primary asset must have a rigorous integrity.
3. Quickly controlled	The primary asset may have an integrity issue if it is discovered and if the integrity is recovered within a short timeframe.
2. Controlled	The primary asset may have an integrity issue if it is discovered and if the integrity is recovered.
1.Detectable	The primary asset may have an integrity issue if it is discovered.
0. No integrity	The primary asset may have an integrity issue even though not discovered.

Table 6: Selected integrity scale

In order to express safety needs in terms of confidentiality, the selected scale is the following

Levels of the scale	Description of the scale
4.Confidential	The primary asset must only be accessible to one or a thin number of entitled persons in restricted situations (ex. access to device IDs).
3.Private	The primary asset must only be accessible to authorised persons on a need-to-know basis in specific situations (ex. persons authorised to access the originating URLs stored in the database).
2.Restricted	The primary asset must only be accessible to authorised persons on a need-to-know basis (ex. persons authorised to access texts stored in the database, which can be of

	a personal nature).
1.Public	The primary asset must be public (ex.: monitoring dashboard results, legal requirements).

Table 7: Selected confidentiality scale

3.3.3.2 Determination of the severity scale

As a reminder, the severity of the impacts of feared events and risks should be determined, where possible, both according to the identifying capacity of a personal data, and according to the harmful character of the impact for individuals.

Impacted individuals are natural persons at large (citizens and internet users, but also end-users whose data may be processed by the system).

In order to determine the severity of feared events and risks, the selected scale is the following:

Levels of the scale	Impact aspects	Description of the scale
4.Critical	Harmful nature	Concerned persons might suffer significant inconvenience, even irreparable, which they might not be able to overcome (financial peril such as important debts or inability to work; long-lasting illness, death...).
	Identifying capacity	It seems extremely easy to identify someone with the concerned data (ex. first name, surname, date of birth and postal address at one given country's population-wide).
3.Important	Harmful nature	Concerned persons might suffer significant inconvenience, which they will be able to overcome, but with serious difficulties (misappropriation of funds, banking ban, damage to property, job loss, law suit, worsening state of health, wide Internet stigmatisation, extinction of a fundamental right due to self-censorship, closing of business webpages...).
	Identifying capacity	It seems relatively easy to identify someone with the concerned data (ex. Twitter pseudonym; or first name, surname and date of birth at one given country's population-wide).
2.Limited	Harmful nature	Concerned persons might suffer significant inconvenience, which they will be able to overcome despite some difficulties (additional costs, refusal of access to commercial services, fear, misunderstanding, stress, minor ailment, temporary extinction of a fundamental right…).
	Identifying capacity	It seems difficult to identify someone with the concerned data but it might happen (ex. first name and surname at one given country's population-wide).
1.Negligible	Harmful nature	The impact on individuals will be weak; inconveniences will be overcome without difficulty (ex.: waste of time in order to perform a given action, irritation...).
	Identifying capacity	It seems nearly impossible to identify someone with the concerned data (ex. first name only at one given country's population-wide).
0.No impact	Harmful nature	There will be no impact on individuals.
	Identifying capacity	Concerned data are not of a personal nature.

Table 8: Selected severity scale

3.3.3.3 Determination of the likelihood scale

The aim of this step is to create a scale of levels of likelihood, to be associated with threat scenarios. This scale should ideally take into account both the ability of the source to act and the vulnerability of the supporting asset. A template that may be used during this step of the assessment is proposed below (with practical examples that might need to be adapted).

In order to determine the likelihood of threat scenarios, the selected scale is the following:

Levels of the scale	Description of the scale
4.Maximal	That certainly will occur / The supporting asset is very vulnerable.
3.High	That should occur / The supporting asset is vulnerable.
2.Significant	That may occur / The supporting asset might be vulnerable.
1.Minimal	That should not occur / There is a very low vulnerability of the supporting asset.

Table 9: Selected likelihood scale

3.3.3.4 Determination of the risk management criteria

The selected risk management criteria are the following

Action	Risk management criteria (rule chosen to carry out the action))
Estimation of feared event	The severity of feared events is estimated by using the scale provided for to that effect. In addition, where security measures or measures aiming to ensure compliance with law are already implemented: ▪ with the highly probable effect of limiting the severity of the feared event, the severity will be estimated a level lower. ▪ with the highly probable effect of eliminating the severity of the feared event, the severity will be estimated at the lowest level.
Evaluation of feared event	The likelihood is estimated taking into account both the ability of the source to act and the vulnerability of the supporting asset, using the scale provided for to that effect. Where several levels of likelihood are identified depending on the threat's source, the maximum value is retained. In addition, where security measures or measures aiming to ensure compliance with law are already implemented: ▪ with the highly probable effect of limiting the likelihood of the feared event, the likelihood will be estimated a level lower. ▪ with the highly probable effect of preventing the feared event to occur, the likelihood will be estimated at the lowest level.
Estimation of risks	The severity of a risk is equal to the severity of the considered feared event. The likelihood of a risk is equal to the maximal likelihood of all threat scenarios that are linked to the considered feared event.
Evaluation of risks	Risks which severity is critical, and risks which severity is important and likelihood high or maximal, are considered as being intolerable. Risks which severity is important and which likelihood is significant, and risks which severity is limited and which likelihood is high or maximal, are considered as being significant. Other risks are considered as being negligible.

Table 10: Selected risk management criteria

3.4 Assessment of the risks to fundamental rights and freedoms

The assessment of the risks to fundamental rights and freedoms implies to study both feared events and threat scenarios before performing a risk analysis.

3.4.1 Study of feared events

The severity of feared events is presented in the following table.

It has to be noted that in order to assess the severity of impacts, there will be no systematic attribution of a value to the identifying capacity of the primary asset[121], for the following reasons:

- The capacity of originating URLs to enable the identification of the author of a publication will most of the time be very high, unless the source changes URL. Therefore, the identifying nature of such data will be considered as potentially critical.
- The identifying capacity of personal data sent by the users of the smartphone app will generally be critical (where they will agree to send their device ID or give contact details in the report title). This identifying capacity is unknown in the other situations (users being not supposed to provide personal information beside their device ID). As a result, the identifying nature of such data will be by default considered as potentially critical.
- The identifying capacity of personal data relating to individuals potentially present in the database may be negligible to critical, depending on the content of the scanned source, and there is no mean to determine it before searching for such a personal data (which is neither the aim of the system, nor an available functionality) or before seeing such personal data inadvertently. Therefore, the identifying nature of such data will be considered as unknown for each primary asset of this kind.
- Primary assets that consist in fundamental freedoms have no identifying capacity.

It has also to be noted that the study of feared events has been performed at the same time as the study of threat scenarios, as the risk analysis and as the identification of risks treatment, for a greater consistency. However, these results are respectively presented in the form of four tables, for a greater clarity.

Finally, as a reminder, this assessment has been performed both without consideration of and taking into account existing measures that aim to control the feared events and to ensure compliance with legal requirements (legal requirements forming part of primary assets in the current PIA). Indeed, both these measures might already be appropriate to lower the likelihood or the severity of the feared event, since they may have the effect of protecting the security needs of primary assets (these measures are mainly prevention and recovery measures), of reducing identified impacts (prediction and preparation, prevention, containment, fight, recovery, restoration, compensation…), of counteracting each identified threat source (prediction and preparation, deterrence, detection, containment…), of ensuring protection against threats (these measures are mainly detection and protection measures), and of reducing supporting assets' vulnerabilities (these measures are mainly prevention and protection measures).

[121] For further explanation see the MANDOLA Deliverable D2.4a (Intermediate) - *Privacy Impact Assessment of the MANDOLA outcomes*, version 2.4a.2 of 11 July 2017, MANDOLA project (Monitoring ANd Detecting OnLine hAte speech) - GA n° JUST/2014/RRAC/AG/HATE/6652, http://mandola-project.eu/.

Analysis of feared events:

Feared event	Security need (value)	Impact (examples)	Severity of the impact on individuals	Existing measures that could reduce the severity	Severity final assessment
Originating URLs (identifying capacity: potentially critical)					
Unavailability	0 (unavailable)	No impact (beyond a right of access almost impossible to exercise - but the lack of URL prevents any accurate identification - seems to be more beneficial to freedoms than their retention).	1 (negligible)	No measure.	1 (negligible)
Integrity compromission	0 (no integrity)	No impact (beyond a right of access almost impossible to exercise - but the lack of URL prevents any accurate identification - seems to be more beneficial to freedoms than their retention).	1 (negligible)	No measure.	1 (negligible)
Confidentiality breach	4 (confidential)	Data misuse (individual's identification).	3 (important)	No measure.	3 (important)
Data relating to the users of the smartphone app (identifying capacity: critical)					
Unavailability	3 (max. 24 hours)	Impossibility to get feedback in relation to a report. Loss of privacy content hosted on the smartphone	2 (limited)	No measure.	2 (limited)
Integrity compromission	3 (quickly controlled)	Impossibility to get feedback in relation to a report.	1 (negligible)	No measure.	1 (negligible)
Confidentiality breach	4 (confidential)	Publication by a malicious person.	3 (important)	No measure.	4 (important)
Data relating to Internet users, including victims and potential perpetrators of hate speech (identifying capacity: unknown)					
Unavailability	0 (unavailable)	No impact.	0 (no impact)	No measure.	0 (no impact)
Integrity compromission	1 (detectable)	No impact for original web or	0 (no impact) for	Hate speech data base security / no measure in	4 (critical)

		social network author; a malicious or accidental modification might lead to identify a new person.	original authors 4 (critical) for potentially new persons	relation with the report storage module.	
Confidentiality breach	4 (confidential)	Publication of the names of authors of hate speech Tweets	3 (important)	Hate speech data base security / no measure in relation with the report storage module.	3 (important)
Compliance with legal requirements					
Unavailability	4 (available)	Data controller missing to comply with his/her legal obligations	4 (critical)	PIA, test of compliance with legal requirement, recommendations of further development and use.	3 (important)
Integrity compromission	4 (integrity)	Data controller complying imperfectly with his/her legal obligations	3 (important)	PIA, test of compliance with legal requirement, recommendations of further development and use.	2 (limited)
Confidentiality breach	0 (public)[122]	Possible impact on certain fundamental freedoms exercise (ex. freedom of speech) if some aspects of the legal compliance are not known.	3 (important)	PIA, test of compliance with legal requirement, recommendations of further development and use.	2 (limited)
Fundamental rights and freedoms[123]					
Unavailability	4 (available)	Arrest following an investigation; loss of a personal account's data (freedom of privacy); self-censorship (expression).	3 to 4 (might be critical in some situations, important in most of other ones)	PIA, test of compliance with ECHR requirement, awareness messages included in the dashboard and smartphone app, care in presentation of results per country/city; securisation of systems remaining under MANDOLA partners' control; recommendations of further development and use.	3 (important)
Integrity compromission	4 (integrity)	Short-term investigation; webpage or Twitter account closure (freedom of trade,	3 (important)	PIA, test of compliance with ECHR requirement, awareness messages included in the dashboard and smartphone app, care in presentation of	2 (limited)

[122] This value is anticipating the introduction of the GDPR. See for example its recital 39: "*Natural persons should be made aware of risks, rules, safeguards and rights in relation to the processing of personal data and how to exercise their rights in relation to such processing*".

[123] Within the limits of their protection by the ECtHR. On this issue see the MANDOLA deliverable D2.2 - *Identification and analysis of the legal and ethical framework*, version 2.2.4 of 12 July 2017, MANDOLA project (Monitoring ANd Detecting OnLine hAte speech) - GA n° JUST/2014/RRAC/AG/HATE/6652, http://mandola-project.eu/.

		presumption of innocence)		results per country/city; securisation of systems remaining under MANDOLA partners' control; recommendations of further development and use.	
Confidentiality breach	4 (confidential)	Publicisation of personal opinions (secrecy of privacy); consequential retaliation or attempts to dignity.	3 (important)	PIA, test of compliance with ECHR requirement, awareness messages included in the dashboard and smartphone app, securisation of systems remaining under MANDOLA partners' control; recommendations of further development and use.	2 (limited)

Table 11: Study of feared events

3.4.2 Study of threat scenarios

Results of the study of threat scenarios are presented in the following table.

It has to be noted that the vulnerability of supporting assets is often considered as significant (as an average taking into account generic security measures), and that this evaluation could have to be reviewed by end-users, since it may depend of elements of contexts that are under their control but that are unknown to the MANDOLA consortium.

Study of threat scenarios

Feared event	**Threat source**	**Probable threats (action)**	**Likelihood**			**Existing measures that could reduce the likelihood**	**Likelihood final assessment**
			Source's ability to act	**Supporting asset vulnerability**	**Total**		
Originating URLs (identifying capacity: potentially critical)							
Unavailability	1. Analyst / IT developer with limited awareness or limited motivation or acting unconsciously; agent of a sub-contractor / hosting provider / services provider with limited awareness or limited motivation or acting unconsciously. 2. Criminal organisation. 3. Malicious code untargeted of unknown origin, computer virus. 4. Lightning, other	1. Data deletion. 2. Acting on the premises (hardware deterioration or modification). 3-6. Hardware or software deterioration or modification.	3 (High)	3 (High)	3 (High)	Security measures applied at FORTH and UCY. Security measures recommended for third parties connected to the app and other data set providers.	2 (significant)

	natural phenomenon, wear. 5. Fire, computer failure. 6. Change in network infrastructure, addition of a new component to the information system.						
Integrity compromission	1. Activist or NGO guided by ideology / politic beliefs; passionate hacker; former agent whishing harming the organisation or pursuing personal interests 2. Malicious code untargeted of unknown origin, computer virus.	1-2. Modification or deterioration of URLs.	3 (High)	3 (High)	3 (High)	Security measures applied at FORTH and UCY. Security measures recommended to third parties connected to the app and other data set providers.	2 (significant)
Confidentiality breach	1. Maintenance/cleaning staff. 2. Analyst with low level of access authorisation; trainee acting playfully or in order to strengthen the combat against online hate speech. 3. Sub-contractor, provider, help-desk agent. 4. Analyst / IT developer with limited awareness or limited motivation or acting unconsciously; agent of a sub-contractor / hosting provider / services provider with limited awareness or limited motivation or acting unconsciously. 5. Head of operations with limited awareness regarding security or personal data protection within the framework of the fight against hate speech 6. Malicious code	1. Accessing paper copies of reports (personal revenge, publication...) 2. Search for the direct or indirect identity of the author of a potentially illegal Tweet --> paper copy or publication. 3-7. Undue access to / collection / reuse of data. 8. Acting on the premises, copying the information.	4 (Maximal)	3 (High)	4 (Maximal)	No particular measures beyond security measures already applied and recommendations to third parties on security matters.	4 (Maximal)

	untargeted of unknown origin, computer virus 7. Activist or NGO guided by ideology / politic beliefs; passionate hacker; former agent whishing harming the organisation or pursuing personal interests. 8. Criminal organisation.						
Data relating to the users of the smartphone app							
Unavailability	1. Staff of an assistance service wishing to prejudice authors of reports. 2. Analyst / IT developer with limited awareness or limited motivation or acting unconsciously; agent of a sub-contractor / hosting provider / services provider with limited awareness or limited motivation or acting unconsciously. 3. Criminal organisation. 4. Person belonging to the user's environment, having access to the smartphone and wishing to protect the user. 5. User him or herself. 6. Future developers of the app. 7. Malicious code untargeted of unknown origin, computer virus 8. Lightning, other natural phenomenon, wear 9. Fire, computer failure, defect of the app. 10. Change in	1. Action on smartphones using the app (data blocking or destruction). 2. Action on servers (destruction of data). 3.Acting on the premises hardware or software destruction or damaging). 4. Action on smartphone (removing the app and -by mistake- the data). 5. Mistake in the software installation or removal. 6. App update - undue removal of personal content. 7. Action on smartphone through the app, on the app, or on the data controllers' servers. 8-10. Hardware or software destruction or damaging.	3 (High)	3 (High)	3 (High)	Security measures applied at FORTH and UCY. Security measures recommended to third parties connected to the app and other data set providers.	2 (servers) 3 (smartphones)

	network infrastructure, addition of a new component to the information system.						
Integrity compromission	1. Developer. 2. Activist or NGO guided by ideology / politic beliefs; passionate hacker; former agent whishing harming the organisation or pursuing personal interests. 3. Malicious code untargeted of unknown origin, computer virus.	1. Creating a weakness in the system in order to later on compromise it. 2. Modifying data. 3. Acting on data controllers' servers or on the app or on the smartphone through the app.	3 (High)	3 (High)	3 (High)	Security measures applied at FORTH and UCY. Security measures recommended to third parties connected to the app and other data set providers.	2 (significant)
Confidentiality breach	1. Sub-contractor, provider, help-desk agent. 2. Person belonging to the user's close environment, knowing the smartphone access code or having the possibility to consult the smartphone, willing to prejudice this user or monitor his/her activity. 3. Developer. 4. Staff of an assistance service wishing to prejudice authors of reports. 5. Analyst / IT developer with limited awareness or limited motivation or acting unconsciously; agent of a sub-contractor / hosting provider / services provider with limited awareness or limited motivation or acting unconsciously. 6. Head of operations with limited awareness regarding security or personal data protection	1. Undue access to / collection / reuse of data. 2. Sending of false reports; access to previous reports in a familiar cell where retaliation is possible - including if previous reports were relating to this family activity. 3. Creating a weakness in the data controller' system in order to later on compromise it. 4. Collection of data on smartphones using the app. 4-6.Collection on the server of personal data including device IDs; display or re-use. 7. Smartphone theft--> password too simple, found --> access to reports. 6. Having kept or accessing devices ID. 8. Compromission of smartphones through the app, or smartphones of which the ID is known (personal data collection, use).	4 (maximal)	4 (maximal)	4 (maximal)	Security measures applied at FORTH and UCY. Security measures recommended to third parties connected to the app and other data set providers.	4 (smartphones) 3 (servers)

	within the framework of the fight against hate speech. 7. Vandal with very common technical skills. 8. Activist or NGO guided by ideology / politic beliefs; passionate hacker; former agent of a reporting platform whishing harming the organisation or the smartphones users or pursuing personal interests. 9. Criminal organisation. 10. Malicious code untargeted of unknown origin, computer virus.	9. Acting on the premises, copying the information. 10. Reaching data controllers' servers or acting on the app/through the app.					
Data relating to Internet users (including victims and potential perpetrators of hate speech)							
Unavailability	1. Criminal organisation. 2. Analyst / IT developer with limited awareness or limited motivation or acting unconsciously; agent of a sub-contractor / hosting provider / services provider with limited awareness or limited motivation or acting unconsciously. 3. Malicious code untargeted of unknown origin, computer virus. 4. Lightning, other natural phenomenon, wear. 5. Fire, computer failure. 6. Change in network infrastructure, addition of a new component to the information system.	1- Acting on the premises, (destruction/damaging systems). 2. Destruction of data. 3-4. Data destruction or software damaging on controllers' servers. 4. Damaging hardware. 5-6. Damaging or disrupting the information system.	3 (High)	3 (High)	3 (High)	Security measures applied at FORTH and UCY. Security measures recommended to third parties connected to the app and other data set providers.	2 (significant)
Integrity compromi	1. Developer. 2. Activist or NGO	1. Creating a weakness in the system in order to later on	3 (High)	3 (High)	3 (Hi	Security measures applied at FORTH and UCY. Security	2 (signific

ssion	guided by ideology / politic beliefs; passionate hacker; former agent whishing harming the organisation or pursuing personal interests. 3. Malicious code untargeted of unknown origin, computer virus.	compromise it. 2. Injection of false data to texts that are stored. 3. Damaging data/softwares on data controllers' servers.			gh)	measures recommended to third parties connected to the app and other data set providers.	ant)
Confidentiality breach	1. Maintenance/cleaning staff. 2. Analyst initiative aiming to strengthen the combat against online hate speech. 3. Activist or NGO guided by ideology / politic beliefs; passionate hacker; former agent whishing harming the organisation or pursuing personal interests. 4. Criminal organisation. 5. Developer. 6. Sub-contractor, provider, help-desk agent. 7. Analyst / IT developer with limited awareness or limited motivation or acting unconsciously; agent of a sub-contractor / hosting provider / services provider with limited awareness or limited motivation or acting unconsciously. 8. Head of operations with limited awareness regarding security or personal data protection within the framework of the fight against hate speech. 9. Malicious code untargeted of	1. Accessing paper copies of reports (and potential further use). 2. Removal of functions protecting users' names and geolocations. 3. Compromising the system (data collection and potential reuse). 4. Acting on the premises, copying the information. 5. Creating a weakness in the system in order to later on compromise it. 5-9. Undue access / collection / reuse of contents. 7. Display of detailed geolocations.	4 (maximal)	4 (maximal)	4 (maximal)	Security measures applied at FORTH and UCY. Security measures recommended to third parties connected to the app and other data set providers; geolocation protection; user names removal.	3 (High)

	unknown origin, computer virus.						
Compliance with legal requirements *(LR impacted are identified in italic and in brackets)*							
Unavailability	1. Head of the organisation/of operations with limited awareness regarding personal data protection or disregarding data protection rules. 2. Developer. 3. Person belonging to the user's close environment, knowing the smartphone access code or having the possibility to consult the smartphone, having high technical skills and willing to prejudice this user or monitor his/her activity. 4. Activist or NGO guided by ideology / politic beliefs; passionate hacker; former agent whishing harming the organisation or pursuing personal interests. 5. Criminal organisation. 6. Internet service provider's personnel wishing to prejudice the user or willing to remove chocking but non-illegal content. 7. Court's decision. 8. Lightning, other natural phenomenon, wear. 9. Fire, computer failure. 10. Change in network infrastructure, addition of a new component to the information system. 11. Sickness.	1. Non implementation or removal of measures aiming at respecting legal requirements, including functions protecting names and geolocations *(potential action on LR1 to LR15).* 2. Creation of a weakness in the system in order to later on compromise it *(potential action on LR2,3,4,5,6,7,9,10,13, 14,15).* 3. Deletion of personal information preventing the exercise of the right of access *(potential action on LR10).* 4. Compromising the system *(potential action on LR2,3,4,5,6,7, 9,10,13,14, 15).* 5. Destruction/damaging systems *(potential action on LR2,3,4,5,6,7,9,10,13, 14, 15).* 6. Closure of the user's account based on the publication of a chocking - alleged illegal - content *(potential action on LR11).* 7. Considering a specific legal basis is required *(potential action on LR1)*. 8-10. Damaging hardware / softwares potentially including technical safeguards (*potential action on LR2,3,4,5,6,7,9,10,13, 14, 15)*. 11. No control of access to personal information; no	4 (maximal)	4 (maximal)	4 (maximal)	Security measures applied at FORTH and UCY. Security measures recommended to third parties connected to the app and other data set providers. MANDOLA recommendations to the Industry. Recommendations concluding the PIA.	3 (High)

		manual data deletion in order to respect time limits; maintenance issues leading to obsolescence or falsehood of the list of reports' recipients, of their DP policies, and of the information rel. to the processing; lack of regulars PIA, no response to data subject access requests, obsolescence of the measures aiming to comply and document compliance with law (*potential action on LR1 to LR15).*					
Integrity compromission	1. Analyst initiative aiming to strengthen the combat against online hate speech, while breaching data protection rules. 2. Sub-contractor, provider, help-desk agent. 3. Analyst / trainee with limited awareness or limited motivation or acting unconsciously. 4. Vandal with very common technical skills. 5. Personnel of one of the data or system controllers.	1. Removal of measures aiming at respecting legal requirements, including functions protecting names and geolocations *(potential action on LR1, LR4 to 7, LR9 to LR15).* 2-3. Undue access to, collection and/or reuse of data *(LR4).* 4. Smartphone theft--> no access to previous reports - might be difficult to exercise access rights *(LR10)*. 5. By negligence or maliciousness: no or partial control of access to personal information; no or partial manual data deletion in order to respect time limits; maintenance issues leading to obsolescence or falsehood of the list of reports' recipients, of their DP policies and of the information rel. to the processing; lack of regulars PIA, no or partial response to data subject access requests, obsolescence	4 (maximal)	4 (maximal)	4 (maximal)	Security measures applied at FORTH and UCY. Security measures recommended to third parties connected to the app and other data set providers. MANDOLA recommendations to the Industry. Recommendations concluding the PIA.	3 (High)

		of the measures aiming to comply and document compliance with law (*potential action on LR1 to LR15).* 5. By negligence, maliciousness or authoritarism: providing wrong advices through the modification of the MANDOLA publications *(LR2, LR3).*					
Confidentiality breach	1. Court decision, Parliament 2. Head of the organisation/of operations with limited awareness regarding personal data protection or disregarding data protection rules. 3. Personnel of one of the data or system controllers by negligence or maliciousness. 4. Sickness. 5. Fire, computer failure. 6. Lightning, other natural phenomenon, wear. 7. Change in network infrastructure, addition of a new component to the information system. 8 Malicious code untargeted of unknown origin, computer virus. 9. Activist or NGO guided by ideology / politic beliefs; passionate hacker; former agent whishing harming the organisation or pursuing personal interests; criminal organisation.	1. Specific legal basis unavailable *(LR1).* 2-4. Non display or obsolescence of measures taken to ensure legal compliance / DPIA results non-publics / non-response to data subjects access requests *(Potential effect on publicity of LR2 to LR15).* 5-9.Damaging information supports (websites, information provided through the app and the dashboard) *(Potential effect on publicity of LR2 to LR15)*. 8-9. Injection of false data to texts that are displayed (websites, information provided through the app and the dashboard) *(Potential effect on publicity of LR2 to LR15).*	4 (maximal)	4 (maximal)	4 (maximal)	Security measures applied at FORTH and UCY. Security measures recommended to third parties connected to the app and other data set providers. MANDOLA recommendations to the Industry. Recommendations concluding the PIA.	3 (High)
Fundamental rights and freedoms							
Unavailability	1. Person belonging to the user's	1. Reporting him/her as author of hate	4 (person	4 (inform	4 to	Security measures applied at FORTH	3

	environment, wishing to prejudice this user. 2. Person belonging to the user's environment, wishing to protect the user. 3. Person belonging to the user's environment, having access to the smartphone and wishing to protect the user. 4. Internet service provider. 5. Activist or NGO guided by ideology / politic beliefs; passionate hacker; former agent of a reporting platform whishing harming the organisation or the smartphones users or pursuing personal interests. 6. Hateful persons or group, including terrorists. 7. User him or herself. 8. LEA. 9. Malicious code untargeted of unknown origin, computer virus. 10. Analyst / trainee with limited awareness or limited motivation or acting unconsciously. 11. Vandal with important technical skills.	speech content *(potential impact on FR1 to FR9).* 1-2. Persuading him or her to not use the app due to the privacy limitation it incurs; persuade him or her to not publish on social networks due to close monitoring. *(potential impact on FR1 to FR3, FR6 and FR9).* 3. Asking a provider to remove a text for privacy reasons and as a result obtaining the closure of the account *(potential impact on FR1 to FR6 and FR9).* 4. Action on the user's account *(potential impact on FR1 to FR6 and FR9).* 5. Compromising smartphones through the app, or smartphones of which the ID is known-> loss of reports and potential other important personal information *(potential impact on FR1, FR2 and FR6)*. 6. Manufacturing false hate speech content + anonymous reports -> investigations against innocent people *(potential impact on FR1 to FR9)*. 7. Self-censorship, mistake in installation/software removal; wrong attitudes due to a bad quality or a bad understanding or an undue modification of the MANDOLA advices. *(potential impact on FR1 to FR4, FR6, FR7, FR9)*. 8. Opening of an investigation based on	s, LEAs, ISPs, user, analyst) 3 (criminals, activists, virus)	ation supports), 3 (systems) 3 (users) 2 (LEA, ISP)	2 depending on the source and support	and UCY. Security measures recommended to third parties connected to the app and other data set providers. MANDOLA recommendations to the Industry and to user; recommendation of objectivity and quality in the drafting of these recommendations. Recommendations concluding the PIA.	

		wrong information; investigating a wrong IP address following a report *(potential impact on FR1 to FR9)*. 9. Coming from web supports of the MANDOLA information *(potential impact on FR1 to FR3, FR7, FR9)*. 10. Assessing contents wrongfully (author targeted as author of a potentially illegal content) *(potential impact on FR1 to FR9)*. 11. Smartphone theft--> password found --> access to reports --> anonymous publication under the user identity *(potential impact on FR1 to FR9)*.					
Integrity compromission	1. Person belonging to the user's close environment, with significant technical skills willing to prejudice this user. 2. Third party without important technical skills. 3. Other user of the Internet service wishing to prejudice someone. 4. Analyst / trainee with limited awareness or limited motivation or acting unconsciously or mislead. 5. Vandal with very common technical skills. 6. Users with poor awareness on legal issues, thinking to have the duty of finding and reporting hate speech. 7. Activist or NGO guided by ideology / politic beliefs; passionate hacker;	1. Publishing false hate speeches under the user's name *(potential impact on FR1 to FR9)*. 2. Sending several times the same report in order to bias results *(potential impact on FR1 to FR9)*. 3. False reports to the MANDOLA system, which can be perceived as hate speech if taken out of their context (ex. theatre scenario) *(potential impact on FR1 to FR9)*. 4. Assessing contents wrongfully (defect of statistics) *(potential impact on FR1 to FR9)*. 5. Smartphone theft--> loss of information of private life (activity relating to reports) *(potential impact on FR1 to FR3)*. 6. Abnormally important number of reports relating to legal contents (may burden stats if not	4 (persons w. technical skills, policy makers) 3 (persons, vandal, criminals, activists, virus).	4 (online information supports, MANDOLA system,) 3 (mandola systems, mandola outcomes, others, users) 2 policy makers	4	Security measures applied at FORTH and UCY. Security measures recommended to third parties connected to the app and other data set providers. MANDOLA recommendations to the users and policy makers; recommendation of objectivity and quality in the drafting of these recommendations. Recommendations concluding the PIA.	3

	former agent whishing harming the organisation or pursuing personal interests. 8. Personnel of one of the data or system controllers. 9. Third party copying the MANDOLA information. 10. Personnel with technical knowledge willing to harm the organisation or the combat against hate or pursuing personal interests. 11. Lightning, other natural phenomenon, wear 12. Sickness 13. Policy makers. 14. Internet users.	appropriately assessed) *(potential impact on FR1 to FR9)*. 7. Modification of MANDOLA publications - providing wrong advices as a result *(potential impact on FR1 to FR9)*. 8-9. By negligence, non-taking into account of legislative modifications relating to illegal hate speech; or maintenance issues leading to obsolescence or falsehood of the dashboard statistics or of the MANDOLA recommendations *(potential impact on FR1 to FR9)*. 8-9-10. Thinking knowing the truth or willing to prejudice MANDOLA outcomes: modification with false information of the MANDOLA information including advices (and 10: statistics), thereby providing false information / wrong advice (Internet users misled on behaviours to be adopted; policy makers misled on decisions to be made) *(potential impact on FR1 to FR9)*. 11-12. Maintenance issues leading to obsolescence or falsehood of the MANDOLA recommendations or dashboard statistics while a regular update was announced (same possible consequences as above) *(potential impact on FR1 to FR9)*. 13-14. Wrong decisions taken on the basis of false information,					

		including due to a misunderstanding of statistics or due to one of the other actions above.					
Confidentiality breach	1. Vandal with very common technical skills.	1. Smartphone theft--> password too simple, found --> access to reports including screenshots in private areas *(potential impact on FR1 to FR3 - data subject: smartphone user and Internet users authors of copied content)*.	3	2	3	No measure beyond those already recommended during the performance of the current PIA step.	3

Table 12: Study of threat scenarios

3.4.3 Risk analysis

The assessment of risks is presented in the following table.

Template that may be used in order to assess risks

Feared event	Threat source	Threat (action)	Severity of the impact column	Likelihood of the threat scenario	Risk assessment
Originating URLs					
Unavailability (Risk 1)	- Analyst / IT developer with limited awareness or limited motivation or acting unconsciously. - Agent of a sub-contractor / hosting provider / services provider with limited awareness or limited motivation or acting unconsciously. - Criminal organisation. - Malicious code untargeted of unknown origin, computer virus. - Lightning, other natural phenomenon, wear. - Fire, computer failure. - Change in network infrastructure, addition of a new component to the information system.	-Hardware or software deterioration or modification. -Data deletion.	1	2	1 (N)
Integrity compromission (Risk 2)	- Activist or NGO guided by ideology / politic beliefs; passionate hacker; former agent whishing harming the organisation or pursuing personal interests. - Malicious code untargeted of unknown origin, computer virus.	- URLs (data) deterioration or modification.	1	2	1 (N)
Confidentiality breach (Risk 3)	- Maintenance/cleaning staff. - Analyst with low level of access authorisation; trainee acting playfully or in order to strengthen the combat against online hate	- Undue access to paper copies. - Undue search for the direct or indirect identity of the author of a potentially illegal Tweet --> paper copy or	3	4	3 (I)

	speech. - Sub-contractor, provider, help-desk agent. - Analyst / IT developer with limited awareness or limited motivation or acting unconsciously; agent of a sub-contractor / hosting provider / services provider with limited awareness or limited motivation or acting unconsciously. - Head of operations with limited awareness regarding security or personal data protection within the framework of the fight against hate speech. - Malicious code untargeted of unknown origin, computer virus. - Activist or NGO guided by ideology / politic beliefs; passionate hacker; former agent whishing harming the organisation or pursuing personal interests. - Criminal organisation.	publication. - Undue access to, collection and re-use of URLs (data).			
Data relating to the users of the smartphone app					
Unavailability (Risk 4)	- Staff of an assistance service wishing to prejudice authors of reports. - Analyst / IT developer with limited awareness or limited motivation or acting unconsciously; agent of a sub-contractor / hosting provider / services provider with limited awareness or limited motivation or acting unconsciously. - Criminal organisation - Person belonging to the user's environment, having access to the smartphone and wishing to protect the user. - User him or herself - Future developers of the app. - Malicious code untargeted of unknown origin, computer virus - Lightning, other natural phenomenon, wear - Fire, computer failure, defect of the app. - Change in network infrastructure, addition of a new component to the information system.	- Action on smartphones using the app (data blocking or destruction). - Action on servers (destruction of data). - Action on servers (hardware or software destruction or damaging). - Accidental deletion of data while installing, updating or removing the app.	2	3	2 (s)
Integrity compromission (Risk 5)	- Developer. - Activist or NGO guided by ideology / politic beliefs; passionate hacker; former agent whishing harming the organisation or pursuing personal interests. - Malicious code untargeted of unknown origin, computer virus.	- Creating a weakness in the system in order to later on compromise it. - Data modification on servers or on the smartphone through the app.	1	2	1 (N)
Confidentiality breach (Risk 6)	- Sub-contractor, provider, help-desk agent. - Person belonging to the user's close environment, knowing the smartphone access code or having the possibility to consult the smartphone, willing to prejudice this user or monitor his/her activity. - Developer. - Staff of an assistance service wishing to prejudice authors of reports. - Analyst / IT developer with limited awareness or limited motivation or acting unconsciously; agent of a sub-contractor / hosting provider / services provider with limited awareness or	- Remote undue access to, collection and re-use of data. - Sending of false reports; - Physical undue access to previous reports (theft + too simple password; or familiar cell where retaliation is possible - including if previous reports were relating to this family activity). - Creating a weakness in the data controller' system in order to later on compromise it. - Undue access (remotely of physically) on hosting servers to data including	4	4	3 (I)

	limited motivation or acting unconsciously. - Head of operations with limited awareness regarding security or personal data protection within the framework of the fight against hate speech. - Vandal with very common technical skills - Activist or NGO guided by ideology / politic beliefs; passionate hacker; former agent of a reporting platform whishing harming the organisation or the smartphones users or pursuing personal interests. - Criminal organisation. - Malicious code untargeted of unknown origin, computer virus.	device IDs; collection or re-use.			
Data relating to Internet users					
Unavailability (Risk 7)	- Criminal organisation. - Analyst / IT developer with limited awareness or limited motivation or acting unconsciously; agent of a sub-contractor / hosting provider / services provider with limited awareness or limited motivation or acting unconsciously. - Malicious code untargeted of unknown origin, computer virus. - Lightning, other natural phenomenon, wear. - Fire, computer failure. - Change in network infrastructure, addition of a new component to the information system.	- Destruction/damaging the data controller information system on premises. - Remote data destruction or software damaging on controllers' servers. - Damaging the data controller's hardware.	0	2	1 (N)
Integrity compromission (Risk 8)	- Developer. - Activist or NGO guided by ideology / politic beliefs; passionate hacker; former agent whishing harming the organisation or pursuing personal interests. - Malicious code untargeted of unknown origin, computer virus.	- Creating a weakness in the system in order to later on compromise it. - Addition of false data to texts that are stored. - Damaging data/softwares on data controllers' servers.	4	2	3 (I)
Confidentiality breach (Risk 9)	- Maintenance/cleaning staff. - Analyst initiative aiming to strengthen the combat against online hate speech. - Activist or NGO guided by ideology / politic beliefs; passionate hacker; former agent whishing harming the organisation or pursuing personal interests. - Criminal organisation. - Developer. - Sub-contractor, provider, help-desk agent. - Analyst / IT developer with limited awareness or limited motivation or acting unconsciously; agent of a sub-contractor / hosting provider / services provider with limited awareness or limited motivation or acting unconsciously. - Head of operations with limited awareness regarding security or personal data protection within the framework of the fight against hate speech. - Malicious code untargeted of unknown origin, computer virus.	- Accessing paper copies of reports (and potential further use). - Removal of functions protecting users' names and geolocations. - Undue data collection and potential reuse (on the premises or remotely). - Creating a weakness in the system in order to later on compromise it. - Display of detailed geolocations.	3	3	3 (I)
Compliance with legal requirements					

Unavailability (Risk 10)	- Head of the organisation/of operations with limited awareness regarding personal data protection or disregarding data protection rules. - Developer. - Person belonging to the user's close environment, knowing the smartphone access code or having the possibility to consult the smartphone, having high technical skills and willing to prejudice this user or monitor his/her activity. - Activist or NGO guided by ideology / politic beliefs; passionate hacker; former agent whishing harming the organisation or pursuing personal interests. - Criminal organisation. - Internet service provider's personnel wishing to prejudice the user or willing to remove chocking but non-illegal content. - Court's decision. - Lightning, other natural phenomenon, wear. - Fire, computer failure. - Change in network infrastructure, addition of a new component to the information system. -Sickness.	- Non implementation or removal of measures aiming at respecting legal requirements, including functions protecting names and geolocations *(potential action on LR1 to LR15).* - Creation of a weakness in the system in order to later on compromise it *(potential action on LR2,3,4,5,6,7,9,10,13, 14,15).* - Deletion of personal information preventing the exercise of the right of access *(potential action on LR10).* - Destruction/damaging systems *(potential action on LR2,3,4,5,6,7,9,10,13, 14, 15).* - Closure of the user's account based on the publication of a chocking - alleged illegal - content *(potential action on LR11).* - Considering a specific legal basis is required *(potential action on LR1).* - Damaging hardware / softwares potentially including technical safeguards (*potential action on LR2,3,4,5,6,7,9,10,13, 14, 15)*. - No control of access to personal information; no manual data deletion in order to respect time limits; maintenance issues leading to obsolescence or falsehood of the list of reports' recipients, of their DP policies, and of the information rel. to the processing; lack of regulars PIA, no response to data subject access requests, obsolescence of the measures aiming to comply and document compliance with law (*potential action on LR1 to LR15).*	3	3	3 (I)
Integrity compromission (Risk 11)	- Analyst initiative aiming to strengthen the combat against online hate speech, while breaching data protection rules. - Sub-contractor, provider, help-desk agent. - Analyst / trainee with limited awareness or limited motivation or acting unconsciously. - Vandal with very common technical skills. - Personnel of one of the data or system controllers.	- Removal of measures aiming at respecting legal requirements, including functions protecting names and geolocations *(potential action on LR1, LR4 to 7, LR9 to LR15).* 2-3. Undue access to, collection and/or reuse of data *(LR4).* - Smartphone theft--> no access to previous reports - might be difficult to exercise access rights *(LR10)*. - By negligence or maliciousness: no or partial control of access to personal information; no or partial manual data deletion in order to respect time limits; maintenance issues leading to obsolescence or falsehood of the list of reports' recipients, of their DP policies and of the information rel. to the processing; lack of regulars PIA, no or partial response to data subject access requests, obsolescence of the measures aiming to comply and document compliance with law (*potential action on*	2	3	2 (S)

		LR1 to LR15). - By negligence, maliciousness or authoritarism: providing wrong advices through the modification of the MANDOLA publications *(LR2, LR3).*			
Confidentiality (publicity) breach (Risk 12)	- Court decision, Parliament - Head of the organisation/of operations with limited awareness regarding personal data protection or disregarding data protection rules. - Personnel of one of the data or system controllers by negligence or maliciousness. - Sickness. - Fire, computer failure. - Lightning, other natural phenomenon, wear. - Change in network infrastructure, addition of a new component to the information system. - Malicious code untargeted of unknown origin, computer virus. - Activist or NGO guided by ideology / politic beliefs; passionate hacker; former agent whishing harming the organisation or pursuing personal interests; criminal organisation.	- Specific legal basis unavailable *(LR1).* - Non display or obsolescence of measures taken to ensure legal compliance / DPIA results non-publics / non-response to data subjects access requests *(Potential effect on publicity of LR2 to LR15).* - Damaging information supports (websites, information provided through the app and the dashboard) *(Potential effect on publicity of LR2 to LR15).* - Injection of false data to texts that are displayed (websites, information provided through the app and the dashboard) *(Potential effect on publicity of LR2 to LR15).*	2	3	2 (S)
Fundamental rights and freedoms					
Unavailability (Risk 13)	- Person belonging to the user's environment, wishing to prejudice this user. - Person belonging to the user's environment, wishing to protect the user. - Person belonging to the user's environment, having access to the smartphone and wishing to protect the user. - Internet service provider. - Activist or NGO guided by ideology / politic beliefs; passionate hacker; former agent of a reporting platform whishing harming the organisation or the smartphones users or pursuing personal interests. - Hateful persons or group, including terrorists. - User him or herself. - LEA. . Malicious code untargeted of unknown origin, computer virus. - Analyst / trainee with limited awareness or limited motivation or acting unconsciously. - Vandal with important technical skills.	- Reporting him/her as author of hate speech content *(potential impact on FR1 to FR9).* - Persuading him or her to not use the app due to the privacy limitation it incurs; persuading him or her to not publish on social networks due to close monitoring. *(potential impact on FR1 to FR3, FR6 and FR9).* - Asking a provider to remove a text for privacy reasons and as a result obtaining the closure of the account *(potential impact on FR1 to FR6 and FR9).* - Action on the user's account *(potential impact on FR1 to FR6 and FR9).* - Compromission of smartphones, loss of reports and potential other important personal information *(potential impact on FR1, FR2 and FR6)*. - Manufacturing false hate speech content + anonymous reports -> investigations against innocent people *(potential impact on FR1 to FR9)*. - Self censorship, mistake in installation/software removal; wrong attitudes due to a bad quality or a bad understanding or an undue modification of the MANDOLA advices or information. *(potential impact on FR1 to FR4, FR6, FR7, FR9)*. - Opening of an investigation based on wrong information; investigating a wrong IP address following a report *(potential impact on FR1 to FR9)*. -. Coming from web supports of the	3	3	3 (I)

		MANDOLA information *(potential impact on FR1 to FR3, FR7, FR9)*. - Assessing contents wrongfully (author targeted as author of a potentially illegal content) *(potential impact on FR1 to FR9)*. - Smartphone theft--> password found --> access to reports --> anonymous publication under the user identity *(potential impact on FR1 to FR9)*.			
Integrity compromission (Risk 14)	- Person belonging to the user's close environment, with significant technical skills willing to prejudice this user. - Third party without important technical skills. - Other user of the Internet service wishing to prejudice someone. - Analyst / trainee with limited awareness or limited motivation or acting unconsciously or mislead. - Vandal with very common technical skills. - Users with poor awareness on legal issues, thinking to have the duty of finding and reporting hate speech. - Activist or NGO guided by ideology / politic beliefs; passionate hacker; former agent whishing harming the organisation or pursuing personal interests. - Personnel of one of the data or system controllers. - Third party copying the MANDOLA information. - Personnel with technical knowledge willing to harm the organisation or the combat against hate or pursuing personal interests. - Lightning, other natural phenomenon, wear - Sickness - Policy makers. - Internet users.	- Publishing false hate speeches under the user's name *(potential impact on FR1 to FR9)*. - Sending several times the same report in order to bias results *(potential impact on FR1 to FR9)*. - False reports to the MANDOLA system, which can be perceived as hate speech if taken out of their context (ex. theatre scenario) *(potential impact on FR1 to FR9)*. - Assessing contents wrongfully (defect of statistics) *(potential impact on FR1 to FR9)*. - Smartphone theft--> loss of information of private life (activity relating to reports) *(potential impact on FR1 to FR3)*. - Abnormally important number of reports relating to legal contents (may burden stats if not appropriately assessed) *(potential impact on FR1 to FR9)*. - Modification of MANDOLA publications providing wrong advices as a result *(potential impact on FR1 to FR9)*. - By negligence, non-taking into account of legislative modifications relating to illegal hate speech; or maintenance issues leading to obsolescence or falsehood of the dashboard statistics or of the MANDOLA recommendations *(potential impact on FR1 to FR9)*. - Thinking knowing the truth or willing to prejudice MANDOLA outcomes: modification with false information of the MANDOLA information including advices (and 10: statistics), thereby providing false information / wrong advice (Internet users misled on behaviours to be adopted; policy makers misled on decisions to be made) *(potential impact on FR1 to FR9)*. - Maintenance issues leading to obsolescence or falsehood of the MANDOLA recommendations or dashboard statistics while a regular update was announced (same possible consequences as above) *(potential impact on FR1 to FR9)*. - Wrong decisions taken on the basis of	2	3	2 (S)

		false information, including due to a misunderstanding of statistics or due to one of the other actions above.			
Confidentiality breach (risk 15)	- Vandal with very common technical skills.	- Smartphone theft--> password too simple, found --> access to reports including screenshots in private areas *(potential impact on FR1 to FR3 - data subject: smartphone user and Internet users authors of copied content)*.	2	3	2 (S)

Legend - (reminder of risk management criteria):

(1) Negligible
(2) Significant
(3) Intolerable

Table 13: Risk analysis

3.4.4 Risk evaluation

Risks analysed above can be evaluated by using to the following table

1 (N) Negligible risks	2 (S) Significant risks	3 (I) Intolerable risks

4 Critical		Risk linked to the integrity compromission of data relating to Internet users **(Risk n°8)**		Risk linked to the confidentiality breach of data relating to the users of the smartphone app **(Risks n°6)**
3 Important			Risk linked to the confidentiality breach of data relating to Internet users **(Risk n°9)** Risk linked to the unavailability of the compliance with one or several legal requirements **(Risk n°10)** Risk linked to the unavailability of fundamental rights protection **(Risk n°13)**	Risk linked to the confidentiality breach of originating URLs **(Risk n°3)**
2 Limited			Risk linked to the unavailability of data relating to the users of the smartphone app **(Risk n°4)** Risk linked to the integrity compromission and to the non-publicity of the compliance with one or several legal requirements **(Risk n°11, Risk n°12)** Risk linked to the integrity compromission of fundamental rights protection **(Risk n°14)** Risk linked to the confidentiality breach of the exercise of fundamental rights protection **(Risk n°15)**	
1 Negligible		Risk linked to the unavailability and integrity compromission of originating URLs **(Risk n°1, Risk n°2)** Risk linked to the integrity compromission of data relating to the users of the smartphone app **(Risk n°5)**		
0 Inexistent		Risk linked to the unavailability of data relating to Internet users **(Risk n°7)**		
Severity / Likelihood	**1 Minimal**	**2 Significant**	**3 High**	**4 Maximal**

Table 14: Risk evaluation

3.5 Risk treatment

Identified risk treatment measures are presented in the following table

Formalisation of risk treatment measures

Description of the measures	Measure's function									Related risks	Measure's nature		
	Prevention (pers. data-PD)	Prevention (legal requirements - LR)	Prevention (fund.rights - FR)	Protection (PD)	Protection (LR)	Protection (FR)	Recovery (PD)	Recovery (LR)	Recovery (FR)		Technical	Organisational	Legal
Implementation by third parties (hotlines and data set providers) and partners -if not already done, security measures in order to avoid external and internal undue access to data, including an authorisation to access personal or potentially personal data (such as Internet texts) on a “need to know” or "need to use" basis in order to perform specific needed task, under agreements of confidentiality and purpose non-diversion. Access control and record of access should be in place as well as a regular independent supervision of past accesses and of their purposes.	x	x	x							1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12	x	x	
Implementation of a functional separation between URLs and relating texts.	x	x	x							3	x		
Implementation of organisational security measures (staff training on basic security behaviours, including securing paper copies and passwords).	x	x	x	x	x	x	x	x	x	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12		x	
To avoid recourse to hosting providers (data sets, hate speech database, report storage modules). Where impossible, to ensure the contractual prohibition of undue access / deletion / modification.	x	x	x							1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12		x	x
To ask to staff and developers the signature of agreements of confidentiality and of non-misuse, recalling sanctions of penal nature (in most countries).	x	x	x							3, 6, 9		x	x

To ensure data regular backup.	x	x	x	x	x	x	x	x	x	6, 10, 11, 12	x		
To ensure the security of the app against external access; To be cautious while updating the app in order to not impact data.	x		x							4,5,6, 9, 10, 11	x		
To ensure the security of the app against undue access of third parties using physically the smartphone, such as enabling the (easy) setting-up of a specific password to access data hosted on the smartphone.	x		x							6, 10, 15	x		
To enable the removal of the app without removing data, or the removal of the data without removing the app; To ask to the user two consecutive positive actions before removing data.	x		x							4, 10	x		
To raise smartphones users' awareness on security issues (theft, passwords, data regular backup, device security updates...).	x		x			x			x	4, 5, 6, 10, 11, 12		x	
To implement security measures ensuring the particular protection of mechanisms designed to ensure the deletion of precise geolocation data and users' names.	x	x	x				x	x	x	9, 10, 11	x		
To designate a data protection officer and to ensure regular legal compliance controls.	x	x	x				x	x	x	10, 11, 12		x	x
To ensure a regular update of the app and of the information it provides.	x	x	x	x	x	x		x	x	4, 5, 6, 10, 11, 12	x		
To propose a point of contact or a help desk in case of issues encountered during the utilisation of the app.	x	x	x		x	x	x	x	x	4, 5, 6	x	x	
Regular awareness on DP and FR issues of the personnel of data and system controllers.	x	x	x	x	x	x		x	x	1 to 15		x	
To perform a new PIA in case of any doubt about the need for additional safeguards including the need for a specific legal basis (for ex. in case of changes in the context of use, of the systems' functions, of the nature of collected data, and of the	x	x	x	x	x	x	x	x	x	1 to 15	x	x	x

processing purposes).													
To ensure that the date of delivery of the MANDOLA recommendations is clearly mentioned, with a specific warning about their possible obsolescence after a certain period of time, including in case of copy and further distribution.			x			x				13, 14		x	
To ensure transparency of all the details of the MANDOLA processing including nature of collected data, sources, purposes, recipients, and safeguards.	x	x	x	x	x	x	x	x	x	13, 14	x	x	x
To recall the necessary neutrality of hosting and access providers towards Internet content, especially based on Directive 2000/31/EC. This includes to recall that they are authorised (if not obliged) to act only in compliance with their domestic law, usually where the content is obviously illegal (the possibility to have effective remedies against ISP's decisions being obligatory), and in a proportionate manner (for example, the closure of an Internet account might not be proportionate in order to remove one single content).			x						x	13			x
To encourage and favour (1) initial and professional LEA training to cybercrime and electronic evidence (*inter alia* in order to ensure LEA knowledge on the possible falsehood of a report, of an Internet content and of a digital direct or indirect identity); (2) a basic awareness of all the judicial personnel (including non-specialised judges and prosecutors) on cybercrime and electronic evidence (*inter alia* in order to ensure a common understanding of current issues and their awareness of existing specialised teams)[124].	x	x	x				x	x	x	13		x	
To recommend to future developers of the MANDOLA			x							14	x	x	

[124] The second part of this recommendation has been added following consultation of the Mandola Advisory Board members.

monitoring dashboard to perform research in order to improve the accuracy of results by taking into account the most possible relevant factors such as those already advised in the recommendation resulting from the analysis of legal and ethical requirements (number of inhabitants, Internet penetration, number of Internet users and frequency and habits in terms of Internet usage; probable competent jurisdiction; context of the speech - which might be a determining factor in the assessment of a content as being potentially illegal - such as cultural aspects , author's intent, polarity, or existence of a public disorder based on the relevant country's courts decisions).													
To recommend to future developers of the MANDOLA monitoring dashboard to perform research in order to enable the non-taking into account of similar reports while calculating the total number of reports received.			x							14	x	x	
To remind to all parties that the MANDOLA outcomes cannot be copied without being accompanied with a reference to their source and their date of publication.			x							14		x	x
To recommend to not authorise any update of the MANDOLA recommendations to users, to policy makers and to the industry before validation by the former MANDOLA partners or an *ad hoc* revision committee offering a guarantee of professionalism. Updates will also have to be accompanied with the update date and the name of updaters.			x							14		x	
To recall the importance to non-take decisions affecting persons on the solely basis of an automated processing (the outcomes of the latter must be corroborated by other information coming from a source of another nature, especially since an electronic identity is easily falsifiable).		x	x							14		x	x

Table 15: Risk treatment

3.6 Stakeholders consultation

This step consists in the consultation of relevant stakeholders who might be impacted by the project or who might bring their expertise to the identification of the risks presented by the project, in order to gather their views. According to the research consortium of the PIAF E.U. project, the objective is to achieve a "'*win-win' result so that everyone benefit*"[125], which is one of the principles of the privacy by design approach.

To this end, the MANDOLA consortium has consulted the members of the project Advisory Board (AB), who belong to several areas of activity (law enforcement, education, industry, civil society combatting illegal online content) and have different but complementary competencies (including technical, legal and ethical).

These stakeholders have been encouraged to share their general opinion and / or to provide for more focus insights, according to their experience, area of expertise and interest, on the following topics:

- The PIA itself, for example its adequacy, its structure, its completeness, and the identification of risks;
- The recommendations which conclude the PIA, including the safeguards proposed in these recommendations (and their appropriateness, their adequacy, a possible lack of safeguard in relation to a particular issue, etc.);
- Any other issue they would like to raise in relation to the current deliverable presenting the PIA (Deliverable D2.4b), its method (Deliverable D2.4a), or the identification of the applicable legal and ethical framework (Deliverable D2.2).

As a result of this consultation, four answers have been received, providing together for six very interesting comments that are summarised below, together with the responses brought to them by the MANDOLA consortium.

1. **Three experts congratulated the work done, considering the PIA recommendations as "well taken" for most of them, or qualifying the work as being "complete" and / or "clearly explained".**

   The MANDOLA consortium has been very pleased to receive this acknowledgement, which reward its efforts in that regard.

2. **According to two experts, the definition of hate speech that was used in the report and during the MANDOLA research, as well as the reason why some offences and not others are included in the definition, were not clear.**

   Indeed, hate speech - and more exactly illegal hate speech - has not been originally defined in the current report since it is the very subject of the MANDOLA Deliverable D2.1 - *Definition of illegal hatred and implications*. As a consequence of this comment, the definition of illegal hate speech that has been used during the MANDOLA research has been clarified in a new Section 3.3.1.3.1 of the current report.

[125] Paul De Hert, Dariusz Kloza, David Wright *et al.*, *Recommendations for a privacy impact assessment framework for the European Union, PIAF (Privacy Impact Assessment Framework) project*, Grant agreement JUST/2010/FRAC/AG/1137 – 30---CE---0377117/00---70, Deliverable D3, November 2012, p. 29, available at http://www.piafproject.eu/Deliverables.html (last accessed on 15 June 2017).

3. **According to one expert, the issues linked with the collection of personal data relating to hate speech victims need also to be taken into account, especially in Section 3.1 (step 3 of the PIA - Determination of the necessity of a PIA and its scale).**

   Indeed, the PIA included all processed personal data in the list of primary assets (in other words in the list of the assets to be protected), but did not focus particularly on victims' personal data. As a consequence of the AB consultation, the PIA has been reviewed in order to take clearly this particular aspect into account. Clarifications have been brought accordingly to Section 3.1, Section 3.3.1.5.1 (Proportionality - Strictly necessary in relation to their scope), Section 3.3.1.5.2 (Legitimate, explicit and specified purpose; Data subject information; Enhanced protection of some sensitive data), Section 3.3.2.1 and Section 3.4.

4. **According to one expert, impacts on fundamental rights of the MANDOLA monitoring dashboard have been accurately assessed but safeguards to be brought must be complemented or more detailed. Indeed, in substance,**

   - **Statistics representing the level of hate speech per city and country must take into account the proportion of inhabitants and of Internet users in each of them, in order to avoid bias.**
   - **Countries must not be considered to present a « dangerous » state of hate, this must be reworded.**
   - **Cultural aspects must be taken into account (for example, hate speech can be culturally trivialised without intent of inciting hate).**
   - **Visible clarifications on the way subjectivity and polarity have been assessed is necessary (in particular, the use of keywords is a limitative methodological shortcut, hate-speech words can be used for other purposes than hate speech and hate speech can exist through metaphors and words shared by some people only).**

   Some of these recommendations had been already made but possibly expressed less clearly or comprehensively. As a result of this comment, Section 3.3.1.5.1 (Proportionality - Strictly necessary in relation to their scope) and related recommendations have been clarified and augmented.

5. **According to one expert, the recommendation to encourage and favour initial and professional LEA training to cybercrime must be accompanied with a similar recommendation targeting all the judicial personnel, in particular non-specialised judges and prosecutors, *inter alia* in order to ensure the awareness of these stakeholders on existing specialised teams.**

   This recommendation has been followed and included in Sections 3.5 and 4.2.6 of the current report.

6. **According to one expert, the summarised recommendations, in the current report, might be difficult to understand for non-legal persons. One solution could be to make links between these recommendations and their justification in previous sections.**

   This recommendation has been followed and has led to the modification of Section 4 of the current report.

The consultation of the members of the MANDOLA Advisory Board has taken place from 11 August to 5 September 2017, and the above-mentioned comments and proposed answers have been presented and discussed at the second Advisory Board meeting held in Brussels on 7 September 2017.

3.7 Monitoring and review

The last step of a PIA is monitoring and review. As explained in the methodology[126], this step aims at subjecting the PIA's results to internal validation, and external audit or review. Ideally, the PIA and its results should be published, primarily in order to enhance trust in the project's results, transparency and accountability of data controllers. In case the PIA highlights that the project "*reveals high residual risks*"[127], the relevant data protection authority must be consulted prior any personal data processing, according to the GDPR and the Directive on personal data protection for the police and criminal justice sector[128], and ideally prior any implementation or use of the project's outcomes.

In addition, the implementation of the measures must be monitored, and mechanisms must ensure that the PIA remains relevant and updated. Especially, the PIA should be revisited each time a modification is brought to the project or processing that has been assessed, the Article 29 data protection working party providing examples of situations which require a new PIA[129]. In addition, a compliance review must be conducted regularly, at the latest three years after the carrying out of the PIA, as advised by the Article 29 data protection working party[130]. A new PIA will be needed, in any case, at the time the future E.U. data protection legal framework will be applicable, in order to ensure compliance with the provisions of this new E.U. legal framework as well as with applicable domestic laws that will be adopted in order to ease and complement the implementation of the General Data Protection Regulation and (where applicable to the assessed project or a part of it) to transpose the Police Directive.

During the MANDOLA project, the monitoring and review task - whose ideal content has been described above, could be performed comprehensively given the particularities of this project.

---

[126] See the MANDOLA Deliverable D2.4a (Intermediate) - *Privacy Impact Assessment of the MANDOLA outcomes*, version 2.4a.2 of 11 July 2017, MANDOLA project (Monitoring ANd Detecting OnLine hAte speech) - GA n° JUST/2014/RRAC/AG/HATE/6652, http://mandola-project.eu/.

[127] Article 29 Data Protection Working Party, *Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679* (WP248), 4 April 2017, p. 17, http://ec.europa.eu/newsroom/document.cfm?doc_id=44137 (last accessed on 15 June 2017).

[128] Art. 36, 1 of the GDPR ; article 26, 1, a of the Directive on personal data protection for the police and criminal justice sector.

[129] Article 29 Data Protection Working Party, *Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679* (WP248), *op. cit.* p. 12. These situations are especially those where there is a "*change to one of the components of the processing operation (data, supporting assets, risk sources, potential impacts, threats, etc.) or (...) (where) the context of the processing evolves (purpose, functionalities, etc.)*". A new PIA is also required, for example, where "*the organisational or societal context for the processing activity has changed, for example because the effects of certain automated decisions have become more significant, new categories of natural persons become vulnerable to discrimination or the data is intended to be transferred to data recipients located in a country which has left the EU*".

[130] Article 29 Data Protection Working Party, *Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679* (WP248), *op. cit.* p. 12.

The current PIA has been twice (before and after consultation of the Advisory Board members) validated internally by all the MANDOLA partners and afterward submitted to an ethical and quality review (two steps that are followed for all MANDOLA deliverables). The deliverable including the PIA of the MANDOLA outcomes has been published and submitted to the European Commission, but the context and framework of the MANDOLA project do not lend themselves well to an additional external audit.

The PIA of the MANDOLA outcomes has been used by the MANDOLA consortium in order to implement privacy by design measures to the utmost possible extent, as well as to issue recommendations of use that target all discovered weaknesses in terms of impact on fundamental rights. However, most of the MANDOLA outcomes are intended to be used by other entities and organisations. As a result, at the end of the MANDOLA project, new PIAs might have to be carried out by these entities and organisations, in order to take into account new contexts of use of these outcomes, and eventually new functions added to technical mechanisms. Monitoring measures and mechanisms designed to ensure that the PIA remains relevant and updated will also have to be designed and implemented by these latter entities and organisations.

4 Summary of recommendations

The following recommendations are divided into two categories: (i) recommendations resulting from the analysis of legal requirements (which have been firstly summarised at the end of Sections 3.3.1.5.1 and 3.3.1.5.2), and (ii) recommendations resulting from the PIA. These recommendations are also the result of the taking into account of the Advisory Board members' opinions summarised in Section 3.6.

4.1 Recommendations resulting from the analysis of legal and ethical requirements (Sections 3.3.1.5.1 and 3.3.1.5.2)

Recommendations below gather the recommendations previously made in this report in relation with the respect of both the ECHR and the E.U. data protection legislation (DPL), and classify them by concerned stakeholder. They are followed by a mark in brackets indicating the legal basis used to identify them (*[ECHR]* or *[DPL]*), and by a footnote indicating the part in Section 3 where the issue has been discussed[131].

4.1.1 Recommendations to the MANDOLA partners (measures implemented during research where not already available)

4.1.1.1 Information of Internet users

- To grant users of the smartphone app with the possibility to not send their smartphone ID to third parties, being warned that, if they make his choice, they will not be informed on the action taken on their report. *[ECHR]*[132]

4.1.1.2 Prevention of discrimination and of arbitrary decisions

Disclaimers in the dashboard

- To include a visible disclaimer in the dashboard pages displaying results, explaining the context of the statistics that are presented (detailing that the dashboard results MIGHT only be illegal, and that this depends on several factors such as their precise legislation and the competent jurisdiction, the context of the content, and the correct assessment of these contents), and the care to be taken when interpreting them. *[ECHR]*[133]
- In addition to this general disclaimer, the dashboard results showing hate-speech by country and city[134] and proposing a hate strength gauge[135] (which enables to obtain a gauge representation of hate strength in specified date range and country) must display a disclaimer that details clearly the variables that are taken into account in order to calculate the hate speech score of countries and cities (such as the number of

[131] These footnotes have been added following consultation of the Mandola Advisory Board members.

[132] See Section 3.3.1.5.1 (proportionality - Strictly necessary in relation to its their scope).

[133] See Section 3.3.1.5.1 (proportionality - Strictly necessary in relation to its their scope).

[134] See the MANDOLA Deliverable D3.1 - *MANDOLA Monitoring Dashboard*, MANDOLA project (Monitoring ANd Detecting OnLine hAte speech) - GA n° JUST/2014/RRAC/AG/HATE/6652, http://mandola-project.eu/, p. 29; pp.35-38.

[135] See the MANDOLA Deliverable D3.1 - *MANDOLA Monitoring Dashboard*, MANDOLA project (Monitoring ANd Detecting OnLine hAte speech) - GA n° JUST/2014/RRAC/AG/HATE/6652, http://mandola-project.eu/, p. 39.

inhabitants and the volume of Internet content produced each day), must avoid the use of the word "dangerous" to qualify countries and cities and must on the opposite explain in simple terms that these statistics cannot represent the state of dangerousness of a given country or city, in particular since (1) they don't take into account several important factors such as the Internet penetration, the number of Internet users and the frequency and their habits in terms of Internet usage; (2) the MANDOLA dashboard shows hate speech that is potentially illegal in one or several E.U. countries but that might not be illegal in one or several others; (3) the context of the speeches are not taken into account and the assessment of contents is not exact science; and (4) even a high level of illegal online hate speeches (which might be produced by the same group of persons, and which are eased by the simplicity of posting on the Internet) does not necessarily means that a given country or city as a whole is dangerous in terms of hate speech usage. *[ECHR]*[136]

- Disclaimers to be implemented in the dashboard results must also be visible when the dashboard results are displayed through the reporting portal and through the smartphone app. *[ECHR]*[137]

Disclaimers in the smartphone app

- At the level of the possibility to send reports from private areas, a disclaimer should warn on the fact that private areas might contain private hate speech, which is not considered to be illegal in most of the studied E.U. countries. *[ECHR]*[138]
- At the level of the possibility to analyse reports in different languages, a disclaimer should warn on the fact that a content that might be illegal or perceived as illegal in one given country might be legal in one other, which means that contents written in different languages might need to be assessed differently. *[ECHR]*[139]

Quality of the information provided to third parties

- Information to be provided by the MANDOLA consortium to Internet users (including a FAQ), policy makers and the industry, through the MANDOLA portal, the MANDOLA reporting portal, the smartphone app and the monitoring dashboard, must be as objective, exhaustive and referenced as possible, with appropriate disclaimers where a given information might encourage behaviours infringing fundamental rights. In particular, it must favour best practices in terms of private initiatives, among those that are respectful for other fundamental rights at stake. This last issue is of importance since online hate speech seems to be accompanied with some actions belonging to private justice[140], which constitute a threat for several fundamental rights and

[136] See Section 3.3.1.5.1 (proportionality - Strictly necessary in relation to its their scope).

[137] See Section 3.3.1.5.1 (proportionality - Strictly necessary in relation to its their scope).

[138] See Section 3.3.1.5.1 (proportionality - Strictly necessary in relation to its their scope).

[139] See Section 3.3.1.5.1 (proportionality - Strictly necessary in relation to its their scope).

[140] MANDOLA Deliverable D4.2 - *Best Practice Guide for Responding to Online Hate Speech for Internet Industry*, March 2017, MANDOLA project (Monitoring ANd Detecting OnLine hAte speech) - GA n° JUST/2014/RRAC/AG/HATE/6652, http://mandola-project.eu/, p. 8; Young People Combating Hate Speech On-line, *Mapping study on projects against hate speech online*, prepared by the British Institute of Human Rights, 15 April 2012, Council of Europe publishing 2012 (DDCP-YD/CHS (2012), https://rm.coe.int/16807023b4 (last accessed on 21 August 2017), Section 2.1.1, 2, p. 9.; MANDOLA Deliverable D2.1b - *Definition of Illegal Hatred and Implications, op. cit.,* Sections 5.4 and 6.2.

freedoms[141]. Disclaimer relating to the FAQ must be visible from the smartphone app and on the MANDOLA reporting portal. *[ECHR]*[142]

- The MANDOLA recommendation of use must make very clear that decisions that might produce legal effects concerning persons potentially identifiable in the database, perpetrators of hate speech offences at the first place, cannot be made on the solely basis of this automated processing (to that purpose this information must be previously corroborated by other information, external to the system). *[DPL]*[143]

4.1.1.3 Anonymisation

- Ideally, no individual should be identifiable in the MANDOLA hate database and, after transmission to relevant LEAs, in the report storage modules. This implies to remove all names and other visible signs that might lead to or that might be personal data. *[DPL]*[144]

4.1.2 Recommendations to future developers of the monitoring dashboard

4.1.2.1 Accuracy of the system's results

- During the subsequent development phases of the dashboard, some research should focus on ways to improve the accuracy of results (while testing this accuracy), by taking into account the most possible relevant factors such as:
  - The number of inhabitants, the Internet penetration, the number of Internet users and the frequency and their habits in terms of Internet usage;
  - The probable competent jurisdiction;
  - The context of the speech such as cultural aspects[145], the author's intent[146], polarity[147], or the existence of a public disorder[148] based on the relevant country's courts decisions. *[ECHR]*[149]

4.1.2.2 Anonymisation

- The mechanism that removes a part of the geolocation coordinates in order to anonymise data collected and shown in the monitoring dashboard is of utmost

[141] See for ex. Council of Europe, *Recommendation CM/Rec(2016)5 of the Committee of Ministers to member States on Internet freedom*, https://search.coe.int/cm/Pages/result_details.aspx?ObjectId=09000016806415fa.

[142] See Section 3.3.1.5.1 (proportionality - Strictly necessary in relation to its their scope).

[143] See Section 3.3.1.5.2 (Prohibition of decisions taken on the solely basis of a data processing).

[144] See Section 3.3.1.5.2 (Legitimate, explicit and specified purpose).

[145] For example, in certain regions, hate speech can be culturally trivialised without intent of inciting hate (culturals aspects and the current footnote have been added following consultation of the Mandola Advisory Board members).

[146] Intention is of high importance and should always be one of the constitutive elements of a hate speech offence. For example, hate speech can be used in the text of a theatre piece in the purpose of denouncing hate. During the MANDOLA Advisory Board consultation, it has been emphasised that hate-speech words can be used for other purposes than hate speech.

[147] In the extension of the previous footnote, one member of the MANDOLA Advisory Board emphasised that hate speech can exist through metaphors and words shared by some people only.

[148] Which might be a requirement, in certain jurisdiction, in order to consider illegal a hate content. For further details see the MANDOLA deliverable 2.1 - *Definition of Illegal Hatred and Implications*, September 2017, MANDOLA project (Monitoring ANd Detecting OnLine hAte speech) - GA n° JUST/2014/RRAC/AG/HATE/6652, http://mandola-project.eu/.

[149] See Section 3.3.1.5.1 (proportionality - Limited by appropriate safeguards (summary)).

importance and must be particularly preserved and secured against removal or circumvention. *[DPL]*[150]

4.1.2.3 Data quality and data subjects' rights of access, communication, rectification and erasure

- It does not seem necessary to recommend that a function is created in order to enable the search, in the hate speech database, for a name of a person in order to ensure data quality and to enable data subjects to exercise their rights of access, communication, rectification and erasure. Indeed, it would favour persons' identification, whereas the system does not pursue this aim (keeping in mind that entities other than LEAs and the judiciary are not entitled to process personal information relating to penal offences).

4.1.3 Recommendations to future developers of the smartphone app

4.1.3.1 Information of Internet users

- To grant smartphone users with the possibility, before the sending of their report, (1) to see the recipient or list of recipients to which their report is intended to be sent, (2) to access the detailed personal data policy of each of these recipients, (3) to remove one or several of these names, and (4) to choose to add one or several names to the list of recipients, at least from a pre-defined list. *[DPL]*[151]
- The above-mentioned detailed personal data policy of each of the recipients will have to be clear and consistent and include from 2018 the GDPR or Police Directive requirements - Articles 13), in addition to include information relating to the purposes of the processing, to the data subject's right of access, communication and erasure in case they decide to send their device ID or other personal information in the report title, and to the contact points to be used in this regard. Ideally (using a privacy by design approach), the latter information should be directly accessible through the smartphone app, and the right of access could also be exercised through the app. *[DPL] [ECHR]*[152]

4.1.3.2 Anonymisation

- To not remove the function of the smartphone app granting users with the possibility to not send their smartphone ID to third parties, being warned that, if they make his choice, they will not be informed on the action taken on their report. *[ECHR]*[153]

4.1.3.3 Data quality and data subjects' rights of access, communication, rectification and erasure

- It does not seem necessary to recommend that a function is created in order to enable the search, in the report storage modules, for a name of a person in order to ensure data quality and to enable data subjects to exercise their rights of access, communication, rectification and erasure. Indeed, it would favour persons' identification, whereas the system does not pursue this aim (keeping in mind that

[150] See Section 3.3.1.5.2 (Legitimate, explicit and specified purpose).

[151] See Section 3.3.1.5.2 (Prohibition of decisions taken on the solely basis of a data processing).

[152] See Section 3.3.1.5.1 (Limited by apropriate safeguards) and Section 3.3.1.5.2 (Data subject information).

[153] See Section 3.3.1.5.1 (proportionality - Strictly necessary in relation to its their scope).

entities other than LEAs and the judiciary are not entitled to process personal information relating to penal offences).

4.1.3.4 Security

- Further developments of the smartphone app should ensure the protection of the content hosted on smartphones. If such a protection was not offered, clear notice should be given to the users in relation to the risks that might be generated by the installation and the use of the software. *[DPL]*[154]
- Device IDs should be stored in a separate database to be accessed only for duly justified reasons, with application of the security measures referred to above. *[DPL]*[155]

4.1.3.5 Protection against data transfer in countries that do not ensure adequate level of protection

- As a precaution in case the list of recipients of the smartphone app would include services that operate in countries were the level of protection might be non-adequate, to warn appropriately the user of the app on the lower state of protection of personal data in certain countries, providing for example a link on the European Commission decisions on the adequacy of the protection of personal data in third countries[156], and advising the user to consult carefully the personal data protection policy of the assistance service that is proposed as recipient of his or her report. *[DPL]*[157]

4.1.4 Recommendations to system or data controllers including third parties connected to the app and MANDOLA partners after the project

4.1.4.1 General legal and ethical compliance

- To respect the data protection legislation beyond these recommendations, which do not intend to be exhaustive but to advise the implementation of safeguards that, in the particular context of the use or further development of the MANDOLA products, ensure completion with law where the latter is too general or might be difficult to apply comprehensively. *[DPL]*[158]
- This includes in particular, where personal data are collected (including indirect ones such as devices ID, Internet texts and URLs), to ensure short time-limitation and security of processing, as well as the provision of detailed personal data policies to be displayed to Internet users. *[ECHR] [DPL]*[159]
- Measures of awareness raising, control and enforcement must also ensure that no modification of the conditions of use of the MANDOLA outcomes and in particular of the MANDOLA technical developments (such as they are described in the current report and

[154] See Section 3.3.1.5.2 (Security and confidentiality of the processing).

[155] See Section 3.3.1.5.2 (Security and confidentiality of the processing).

[156] http://ec.europa.eu/justice/data-protection/international-transfers/adequacy/index_en.htm.

[157] See Section 3.3.1.5.2 (Adequate level of protection in some case of data transfers).

[158] See Section 3.3.1.5.1 (Legal basis).

[159] See Section 3.3.1.5.1 (Strictly necessary in relation to its their scope; Limited by appropriate safeguards) and Section 3.3.1.5.2 (Time limitation ; Security and confidentiality of the processing ; Data subject information).

other MANDOLA technical deliverables) are tolerated, notably at the occasion of the integration of another component to the systems, or/and by using these systems in order to identify individuals or in order to process voluntarily personal data, without performing a new identification of the appropriate safeguards that must be implemented, since the conclusions of the current PIA as a whole and of each of its steps would not be adapted anymore to such a new situation. *[ECHR] [DPL]*[160]

- As a result, any modification of purposes or adjunction of technical functions imply the performance of a specific PIA in order to identify the possibility to pursue the new purposes or to implement the new functions, and, if so, in order to identify the new appropriate safeguards that are needed in this regard (including a specific legal basis, particularly if the system is intended to be used by public authorities or law enforcement services). Regular PIAs will *inter alia* need to ensure that statistics stay up-to-date and consider potential legislative changes, and that the reporting systems stay user-friendly despite modifications of the technical environment. *[DPL] [ECHR]*[161]
- In any case further technical developments in the pursuit of the MANDOLA purposes must remain confined to the freedoms' limitation that are strictly necessary to these purposes (in order to favour a best identification of potential illegal speeches, a quicker report of these speeches, a better understanding of the phenomenon and to favour best practices in terms of private initiatives), and recommendations and advices that are provided must take into account all the schools of thought in the combat against hate speech area. *[ECHR]*[162]
- Regular PIAs of the technical outcomes will also have to be conducted in order to verify that results of the former one are still valid. *[DPL] [ECHR]*[163]
- Regular PIAs (which might be small-scale ones[164]) must be performed in relation to the broadcasting and use of the MANDOLA information provided to policy makers, to the industry and to Internet users, particularly where this information is susceptible to be obsolete or if assets of sciences that have been considered in order to write this information are susceptible to have evolved. This, unless a disclaimer sheds clearly light on the information's date of production, and warns on a risk of obsolescence after a certain period of time. *[ECHR]* [165]

4.1.4.2 Data protection authorities' supervision

- Consultations with relevant supervisory authorities prior processing might be requested from 2018. Indeed this obligation will apply in case the processing would result in a high

[160] See Section 3.3.1.5.1 (Limited by apropriate safeguards) and Section 3.3.1.5.2 (Legitimate, explicit and specified purpose).

[161] See Section 3.3.1.5.1 (Limited by apropriate safeguards) and Section 3.3.1.5.2 (Legitimate, explicit and specified purpose).

[162] See Section 3.3.1.5.1 (Limited by apropriate safeguards).

[163] See Section 3.3.1.5.1 (Limited by apropriate safeguards) and Section 3.3.1.5.2 (Legitimate, explicit and specified purpose).

[164] See the MANDOLA Deliverable D2.4a (Intermediate) - *Privacy Impact Assessment of the MANDOLA outcomes*, version 2.4a.2 of 11 July 2017, MANDOLA project (Monitoring ANd Detecting OnLine hAte speech) - GA n° JUST/2014/RRAC/AG/HATE/6652, http://mandola-project.eu/, Section 4.1.

[165] See Section 3.3.1.5.1 (Limited by apropriate safeguards).

risk in the absence of measures taken by the controller to mitigate the risk, and it appears that high risks could result - *inter alia* - from an intrusion in the smartphone app hosted on users' devices, and from a diversion or an extension of purposes and technical functions of the dashboard. *[DPL]*[166]

4.1.4.3 Information of Internet users

- Once the monitoring dashboard and the smartphone app will be operational, a clear, consistent and visible information should be provided in relation to the purposes of these technical mechanisms, to the data controllers and contact details of data protection officers where applicable, to the processing operations and purposes which are authorised to third parties that will receive reports through the app, to the exact nature of potentially personal data that may be included in databases, to data sources, to their right of access, communication and erasure and the contact points to be used in this regard, and to measures put in place in order to ensure the protection of privacy and personal data (including confidentiality, security and deletion). *[DPL]*[167]
- Ideally, this information should be available in all supports of the MANDOLA outcomes, including - where applicable - the MANDOLA website, the MANDOLA reporting portal, information provided through the smartphone app and the MANDOLA dashboard. It should also be included in all the supports and channels that will give access to the MANDOLA dashboard results. *[DPL]*[168]

4.1.4.4 Prevention of discrimination and of arbitrary decisions

- Persons who access the hate speech database and the report storage module should also see disclaimers highlighting the potential unreliability of hosted data due to the nature of the system, to the nature of information sources, and to the nature of the information itself, in addition to the prohibition to use the content of the database in order to identify a particular person. *[DPL]*[169]
- The fact that a name of person or a sign/a sentence that might identify a person, included in the database, will never correspond for sure to a real and identifiable natural person, must be very clear for any user of the MANDOLA hate speech database and of the report storage module. *[DPL]*[170]

4.1.4.5 Anonymisation

- Once the smartphone owner has received feedback from the recipient of his or her report, his or her device ID should be deleted, as well as any personal information included in the title of the report. *[DPL]*[171]

---

[166] See Section 3.3.1.5.2 (Data protection authority supervision).

[167] See Section 3.3.1.5.2 (Data subject information).

[168] See Section 3.3.1.5.2 (Legitimate, explicit and specified purpose).

[169] See Section 3.3.1.5.2 (Data quality).

[170] See Section 3.3.1.5.2 (Legitimate, explicit and specified purpose).

[171] See Section 3.3.1.5.2 (Data minimisation).

- After the processing of reports and their transmission to the relevant law enforcement services and eventually other legitimate partners[172], reported contents that might contain personal data and URLs related to stored texts should be deleted by the controllers of the report storage modules. *[DPL]*[173]

4.1.4.6 Time limitation

- Deletion policies should be implemented by operators of data sets aiming at training the hate speech classifier, and by third parties receiving reports in the report storage module. These policies should organise the regular deletion of URLs and of all the texts that might contain indirect personal data, as long as they are not absolutely useful to the proper functioning of the system. These policies should be accompanied with procedural measures ensuring that time limits are observed, and subject to periodic review of the need for the storage of data, in order to ensure it fits with the evolution of the processing's context. Data that are blocked instead of erased should only be processed for the purpose which prevented their erasure, and by specially authorised persons. *[DPL]*[174]

4.1.4.7 Security

- Since simple texts might occasionally lead to identify an individual, even if all (technically) visible personal data are removed, as well as URLs each time they are kept, appropriate technical and organisational measures against undue internal or external access must be applied in order to ensure that (1) access to a text or to a URL stored in the hate speech database pursues the solely aim of verifying the illegal nature of one given content, in order to enhance the performances of the dashboard (where applicable), and to ensure that (2) access to a text or an URL stored in a report storage module pursues the solely aim of assessing the illegal nature of the content, in compliance with domestic law and the policies of the assistance service, and forward it to competent authorities. *[DPL]*[175]
- To this end technical (to be implemented by future developers of the hate speech database and of the report storage module) and organisational (to be implemented by data controllers) measures should ensure that access to contents and to URLs is restricted to identified persons accredited to do it on a "need to know" or "need to use" basis in order to perform specific needed tasks (such as maintenance or hate speech validation), under agreements of confidentiality and purpose non-diversion. Access control and record of access should be in place as well as a regular independent supervision of past accesses and of their purposes. *[DPL]*[176]

[172] For instance, assistance services part of the INHOPE network use (where law authorises it) to also send the report to the relevant assistant service, if the content is hosted in its territory, and to the hosting provider, where law imposes an action from the latter.

[173] See Section 3.3.1.5.2 (Data minimisation).

[174] See Section 3.3.1.5.2 (Time limitation).

[175] See Section 3.3.1.5.2 (Security and confidentiality of the processing).

[176] See Section 3.3.1.5.2 (Legitimate, explicit and specified purpose).

- Recourse to hosting providers should be avoided, and where impossible strong contractual and security measures should prevent any undue access, modification, record and other processing of data by a hosting provider or technical provider which services would be used by the operators of the hate speech database or of the report storage module. *[DPL]*[177]

4.1.5 Recommendations to LEA, policy makers and States

4.1.5.1 Prevention of discrimination and of arbitrary decisions

- Strategic decisions taken on the basis of the MANDOLA monitoring dashboard should not restrict some individual's rights. If such a restriction of rights is scheduled, these decisions should not be taken if not corroborated by additional information obtained from another source than the MANDOLA system. *[DPL]*[178]

4.2 Recommendations resulting from the risk treatment analysis (Section 3.5)

4.2.1 Recommendations to the MANDOLA partners (measures implemented during research where not already available)[179]

- To implement security measures in order to avoid external and internal undue access to or a modification of the information remaining under the control of one or several MANDOLA partners, namely the hate speech database and the reporting portal. These measures must include, where applicable, the possibility for authorised personnel only to access personal or potentially personal data (such as Internet texts) on a "need to know" or "need to use" basis in order to perform specific needed tasks, under agreements of confidentiality and purpose non-diversion. Access control and record of access should be in place as well as a regular independent supervision of past accesses and of their purposes.
- To implement organisational security measures, where applicable, such as staff training on basic security behaviours, including securing paper copies and passwords.
- To ensure regular awareness on data and fundamental rights protection issues of their personnel, where applicable.
- To avoid recourse to hosting providers. Where impossible, to ensure the contractual prohibition of undue access / deletion / modification.
- To ask to staff and developers the signature of agreements of confidentiality and of non-misuse, recalling sanctions of penal nature (in most countries).
- To ensure data regular backup.

[177] See Section 3.3.1.5.2 (Legitimate, explicit and specified purpose).

[178] See Section 3.3.1.5.2 (Prohibition of decisions taken on the solely basis of a data processing).

[179] In order to identify the risk or risks addressed by each of the following recommendations, please refer to the table available in Section 3.5 of the current report.

- Where even potentially personal data are processed, to designate a data protection officer and to ensure regular legal compliance controls.
- To perform a new PIA in case of any doubt about the need for additional safeguards including the need for a specific legal basis (for ex. in case of changes in the context of use, of the systems' functions, of the nature of collected data, and of the processing purposes).

4.2.2 Recommendations to the MANDOLA consortium (measures implemented during research) and to future broadcaster of the MANDOLA information and technical developments

- To ensure that the date of delivery of MANDOLA recommendations is clearly mentioned, with a specific warning about their possible obsolescence after a certain period of time, including in case of copy and further distribution.
- To ensure transparency of all the details of the MANDOLA processing including nature of collected data, sources, purposes, recipients, and safeguards.
- To remind to all parties that the MANDOLA outcomes cannot be copied without being accompanied with a reference to their source and their date of publication.
- In case of copy or distribution of the MANDOLA outcomes, to clearly mention that these latter might become obsolete after a certain period of time.
- To not authorise any update of the MANDOLA recommendations to users, to policy makers and to the industry before validation by the former MANDOLA partners or by an *ad hoc* revision committee offering a guarantee of professionalism. Updates will also have to be accompanied with the update date and the name of updaters.

4.2.3 Recommendations to future developers of the monitoring dashboard

- To implement a functional separation between URLs and relating texts.
- To implement security measures ensuring the particular protection of mechanisms designed to ensure the deletion of precise geolocation data and users' names.
- To perform research in order to improve the accuracy of results by taking into account the most possible relevant factors such as those already advised in the recommendation resulting from the analysis of legal and ethical requirements (number of inhabitants, Internet penetration, number of Internet users and frequency and habits in terms of Internet usage; probable competent jurisdiction; context of the speech - which might be a determining factor in the assessment of a content as being potentially illegal - such as cultural aspects , author's intent, polarity, or existence of a public disorder based on the relevant country's courts decisions).
- To perform research in order to enable the non-taking into account of similar reports while calculating the total number of reports received.
- On the opposite, the implementation of a functionality designed to enable the search, in the hate speech database, for particular words, which could ease data subjects' right of access, is not desirable since it would lead to increase the risk of accidental

identification of authors of potentially illegal hate speeches. Other recommendations seem sufficient in order to ensure the protection of these authors' fundamental rights.

4.2.4 Recommendations to future developers of the smartphone app

- To ensure the security of the app against external access.
- To be cautious while updating the app in order to not impact data.
- To ensure the security of the app against undue access of third parties using physically the smartphone, such as enabling the (easy) setting-up of a specific password to access data hosted on the smartphone.
- To ensure a possibility to remove the app without removing data or to remove data without removing the app; to ask from the user two consecutive positive actions before removing data.
- To raise smartphones users' awareness on security issues (theft, passwords, data regular backup, device security updates...).

4.2.5 Recommendations to system or data controllers (including third parties connected to the app)

- To implement security measures in order to avoid external and internal undue access to data, including the possibility for authorised personnel only to access personal or potentially personal data (such as Internet texts, URLs or device IDs) on a "need to know" or "need to use" basis in order to perform specific needed tasks, under agreements of confidentiality and purpose non-diversion. Access control and record of access should be in place as well as a regular independent supervision of past accesses and of their purposes.
- To implement organisational security measures such as staff training on basic security behaviours, including securing paper copies and passwords.
- To ensure regular awareness on data and fundamental rights protection issues of their personnel.
- To avoid recourse to hosting providers. Where impossible, to ensure the contractual prohibition of undue access / deletion / modification.
- To ask to staff and developers the signature of agreements of confidentiality and of non-misuse, recalling sanctions of penal nature (in most countries).
- To ensure data regular backup.
- To designate a data protection officer and to ensure regular legal compliance controls.
- To ensure a regular update of the app and of the information it provides.
- To propose a point of contact or a help desk in case of issues encountered during the utilisation of the app.
- To perform a new PIA in case of any doubt about the need for additional safeguards including the need for a specific legal basis (for ex. in case of changes in the context of use, of the systems' functions, of the nature of collected data, and of the processing purposes).

4.2.6 Recommendations to LEA, policy makers and States

- To encourage and favour initial and professional LEA training to cybercrime and electronic evidence (*inter alia* in order to ensure LEA knowledge on the possible falsehood of a report, of Internet content and of a digital direct or indirect identity).
- To encourage and favour a basic awareness of all the judicial personnel (including non-specialised judges and prosecutors) on cybercrime and electronic evidence (*inter alia* in order to ensure a common understanding of current issues and the awareness of these stakeholders on existing specialised teams)[180].
- To keep in mind the importance to non-take decisions affecting persons on the solely basis of an automated processing (the outcomes of the latter must be corroborated by another information coming from a source of another nature, especially since an electronic identity is easily falsifiable).

4.2.7 Recommendations to all stakeholders

- To keep in mind the necessary neutrality of hosting and access providers towards Internet content, especially based on Directive 2000/31/EC. This includes that they are authorised (if not obliged) to act in compliance with their domestic law only, usually where the content is obviously illegal (the possibility to have effective remedies against ISP's decisions being obligatory), and in a proportionate manner (for example, the closure of an Internet account might not be proportionate in order to remove one single content).

[180] This recommendation has been added following consultation of the Mandola Advisory Board members.

5 Conclusion

The PIA of the MANDOLA outcomes, enriched with the contribution of the members of the MANDOLA Advisory Board, demonstrates that the MANDOLA products are necessary, proportionate and do respect fundamental rights at stake, including the right to private life, to freedom of expression and to personal data protection, provided that recommendations summarised in Section 4 of the current report are implemented and effective.

6 List of main acronyms and abbreviations

AB: MANDOLA Advisory Board

DPIA: data protection impact assessment.

DPL: data protection legislation.

ECHR: European Convention on Human Rights.

ECtHR: European Court of Human Rights.

FR: fundamental rights.

GPDR or "General Data Protection Regulation": refers to Regulation (EU) 2016/679 of the European Parliament and of the Council on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC[181].

LEA: law enforcement authorities.

LE agent / officer / organisation: law enforcement agent / officer / organisation.

LR: legal requirements.

PD: personal data.

PIA: privacy impact assessment.

Police Directive or "Directive on personal data protection for the police and criminal justice sector": refers to the Directive (EU) 2016/680 of the European Parliament and of the Council on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA[182].

[181] http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2016.119.01.0001.01.ENG&toc=OJ:L:2016:119:FULL (last accessed on 12 May 2017).

[182] http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2016.119.01.0089.01.ENG&toc=OJ:L:2016:119:FULL (last accessed on 12 May 2017).

7 Members of the MANDOLA Advisory Board who contributed to the privacy impact assessment

- **Mr. Christian Aghroum**,
  CEO of SoCoA Sàrl (www.socoa.ch),
  Vice President of CyAN (www.cyan.network).

- **Mrs. Fabienne Baider**,
  Associate Professor, Department of French and European Studies, University of Cyprus.

- **Mr. Philippe Jougleux**,
  Associate Professor, Law School, European University Cyprus.

- **Mr. Jean-Christophe Le Toquin**,
  President, CyAN Network.