PROTOCOL TO COMBAT ILLEGAL HATE SPEECH ONLINE

MINISTERIO
DE INCLUSIÓN, SEGURIDAD SOCIAL
Y MIGRACIONES

SECRETARÍA DE ESTADO
DE MIGRACIONES

oberaxe
Observatorio
Español
del Racismo
y la Xenofobia

Co-funded by
the European Union

Catalogue of publications of the General State Administration

https://cpage.mpr.gob.es

Edited and distributed by: Spanish Observatory on Racism and Xenophobia
José Abascal, 39, 28003 Madrid
Email: oberaxe@inclusion.gob.es
Website: http://www.inclusion.gob.es/oberaxe/es/index.htm

NIPO PDF: 121-22-006-9

Design and layout: Carmen de Hijes

PROTOCOL TO COMBAT ILLEGAL HATE SPEECH ONLINE

Foreword

Representatives of the General Council of the Judiciary, the Office of the Prosecutor General, the Secretariats of State for Justice, Security, Education, Sports, Equality, Social Rights and Migration and the Centre for Legal Studies; representatives of the Social Forum for Immigrant Integration, of the State Council of the Roma People, of the Council of Victims of Hate Crimes and Discrimination, of the Spanish Federation of Lesbians, Gays, Transsexuals and Bisexuals, of the Platform of Children's Organisations and of the Third Sector Platform; and the Spanish Digital Economy Association which includes hosting service providers such as YouTube, Facebook, Instagram, Twitter and Microsoft, believe that the Internet makes a positive contribution to innovation, economic growth and communication between citizens and facilitates public debate and the exchange of information, opinions and ideas.

However, concerned about the spread of illegal hate speech on the Internet that threatens the individuals and groups it targets and negatively impacts those who stand up for freedom and tolerance and that challenges democratic speech and harmonious interaction, and surmising that in the current context of the health, economic and social crisis caused by the COVID-19 pandemic, hate speech may intensify.

In accordance with legislation that guarantees the right to freedom of expression and information, they have drawn up this "**Protocol to combat illegal hate speech online**" (hereinafter *the Protocol*) as an instrument that facilitates effective collaboration among the actors involved in combating illegal hate speech online in Spain: institutions of the public administration, civil society organizations and hosting service providers.

Internet hosting service providers play an important role in combating illegal content spread online and in supporting training and awareness-raising of citizens, without prejudice to meeting their specific legal and social responsibilities ensuring the right to freedom of expression.

In addition, and no less importantly, numerous civil society organizations and associations, especially those recognized as trusted flaggers, contribute to combating hate speech online by monitoring Internet content, preparing and disseminating counter-narratives, training hate speech 'activists' and reporting illegal content.

The goal of the *Protocol* is to define and facilitate collaboration between and among all the signatories, each within the scope of its capabilities and remit, in combating illegal hate speech online, focusing on the specific situation in Spain and applying the pertinent Spanish national legislation.

The *Protocol* has therefore been devised as a cooperation and coordination mechanism between Spain's national authorities entrusted with enforcing legislation prohibiting online hate crime and those authorities that combat illegal online hate speech outside the sphere of criminal law. It also seeks to ensure coordination with civil society organizations and Internet hosting service providers.

With regard to the Judiciary, the *Protocol* establishes indicative or representative criteria to potentially constitute a frame of reference. However, in accordance with legislation that guarantees the right to freedom of expression and information, such a reference would neither impact nor compromise the exercise of jurisdiction. Thus, it cannot interfere with or hinder action taken by judicial authorities in adopting measures to restrict information society services to prevent the ongoing dissemination of certain services or illicit content under the provisions of applicable Spanish law.

The *Protocol* will remain open to new members and to future reviews of its scope.

The *Protocol* is based on the "Code of Conduct on countering illegal hate speech online" signed by the European Commission and several hosting service providers in 2016, and on Commission Recommendation (EU) 2018/334 of 1 March 2018 on measures to effectively tackle illegal content online (hereinafter, the Recommendation).

The terminology used in the original Spanish version of the *Protocol* matches the definitions found in Chapter 1 of the official Spanish version of the aforementioned EU Recommendation 2018/334.

The *Protocol* is divided into 6 sections. **Section I** defines hate speech crimes under national Spanish law. It also lists the main European and international standards used to assess the concept of hate speech crimes. Procedures for notifying, communicating, removing or blocking illegal online content are described, as are safeguards under Spanish legislation for hosting service providers when they take action on their own initiative to identify, remove, block or restrict access to illegal content or content that violates their own standards or terms of service. It also describes when and how content providers that send potentially illegal content to a data hosting platform must be informed.

Section II proposes preparing a list of competent authorities that shall take responsibility for reporting illegal hate speech online. Following the recommendation of the European Commission to establish a Point of Contact for competent authorities to communicate with hosting service providers, the Computer Crime Unit of the Office of the Prosecutor General has been appointed for that purpose. The Point of Contact will facilitate notification of illegal content through a straightforward procedure providing safeguards to block, remove or restrict access to this content, thus contributing to the effective enforcement of Spanish law. It describes how the Point of Contact may require the hosting service provider to refrain from informing the data provider of the removal or blocking of content when such content constitutes a serious crime. It also proposes preparing a form for the competent authorities to use for notification purposes.

Section III proposes that hosting service providers preferentially process notices from trusted flaggers duly accredited as such. It is also proposes preparing a form to file such notices and their content.

Section IV focuses on the accreditation and training of trusted flaggers. It provides for the creation of a Trusted Flagger Accreditation Committee, its composition and the selection of trusted flaggers. It

proposes that the government administration and hosting service providers train trusted flaggers to ensure that they are familiar with the rules governing the use of hosting platforms and Spanish law on illegal hate speech.

Section V addresses the implementation of redress mechanisms which is also provided for under EU Recommendation 2018/334. The aim here is to provide citizens with information on alternative hate speech dispute settlement mechanisms without having to turn to the criminal courts to enforce applicable Spanish law.

Section VI focuses on the implementation and monitoring of the *Protocol* which involves the Monitoring Committee under the *Inter-institutional Agreement*. This Committee will create a collaboration mechanism engaging signatories of the *Protocol* in its implementation. Moreover, activity reports will be prepared and sent to the *Inter-institutional Agreement*'s Monitoring Committee.

Protocol

I. UNDERSTANDING 'ILLEGAL HATE SPEECH' AS 'ILLEGAL CONTENT' UNDER THE PROTOCOL

I.1 For the purposes of this Protocol, it is assumed that *illegal hate speech* refers to *hate speech crime*, i.e. the behaviours described under Article 510 of the Criminal Code or the crimes described under Spanish criminal law consisting of acts of expression-communication to which Article 22(4) of the Criminal Code apply, and to hate speech that may be included among the offences set forth in sections b) and c) of Article 23(1) of Law 19/2007 of 11 July 2007 prohibiting violence, racism, xenophobia and intolerance in sports, provided that these acts on the Internet have led to the hosting of content by hosting service providers. In assessing the concept of illegal hate speech, due consideration will be given to Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law; Council of Europe Recommendation No. R (97) 20 of the Committee of Ministers; ECRI General Policy Recommendation No. 15 of 2016; and General Recommendation No. 35 on combating racist hate speech of the United Nations Committee on the Elimination of Racial Discrimination, provided that the allegedly racist content can be qualified as infringing the aforementioned Spanish laws.

This notwithstanding, hosting service providers shall also assess notices and communications in accordance with their own policies, terms of service and community norms and standards.

All of this in full compliance with the Charter of Fundamental Rights of the European Union, in particular the right to freedom of expression and information, and other applicable provisions of EU law, especially

concerning protection of personal data, competition and electronic commerce. It will also be taken into consideration that, as the European Court of Human Rights has concluded, freedom of expression protects not only opinions and ideas that are favourably received or considered inoffensive or of no account, but also opinions and ideas that may offend, shock or disturb the State or any sector of the population. This is a prerequisite for pluralism, tolerance and a spirit of openness without which there would be no democratic society.

I.2 When hosting service providers receive notice of content that may be considered illegal hate speech, they shall assess the evidence of illegal hate speech set forth in the notice or communication to determine the appropriateness of blocking, removing, restricting access, etc. to said content.

This Protocol shall apply without prejudice to the position of hosting service providers in accordance with Directive 2000/31/EC and Law 34/2002 of 11 July 2002 on Information Society and Electronic Commerce Services (LSSI) or the equivalent Directive and Law that may be in force in the future. Hosting service providers shall not be held responsible for the information that they store, index, make available or transmit if they simply and in good faith, in compliance with this Protocol or, on their own initiative, take voluntary, automated or non-automated action to identify, remove, block or restrict access to illegal content, or if the hosting service providers consider that said information violates their own policies, terms of service, standards or community norms.

Specifically, when hosting service providers undertake a voluntary action, it shall not be assumed that they have knowledge or control of the information they transmit or store, nor shall it be assumed that the activity of the hosting service providers has ceased to be anything more than technical, automatic and passive. The foregoing shall not interfere with a competent authority's ability to order the hosting service provider to put an end to or prevent a specific infringement.

II. NOTICES FROM THE 'COMPETENT AUTHORITIES' AND 'POINT OF CONTACT'

II.1 A limited number of competent authorities, that must be government entities, shall be established for the purposes of the *Protocol* and shall be communicated to hosting service providers.

II.2 Without prejudice to those who are or will be on this list of competent authorities, 'notices from the competent authorities' in the terms defined by the Recommendation shall be restricted to those made through the Point of Contact. To that end, a dual system is proposed: a Point of Contact (through which only notices of illegal content from the competent authorities are channelled) and a list of competent authorities (only relevant within the territory of a Member State) responsible for forwarding information of this nature to the Point of Contact.

II.3 The Point of Contact shall be the Computer Crime Unit of the Office of the Prosecutor General since, in accordance with the principles of unity of action and organisational structure governing the actions of the Public Prosecutor's Office and pre-existing regulations and action protocols, the Office of the Prosecutor General already channels or could channel notices to the Point of Contact. Through its specialized police units, the Ministry of the Interior undertakes to provide the necessary assistance to the Computer Crime Unit of the Office of the Prosecutor General within the framework of this general action protocol.

In any event, action taken by the Public Prosecutor's Office is constitutionally subject to the principles of legality, impartiality, unity of action and hierarchy and therefore cannot be conditioned by the agreements adopted within the scope of this *Protocol*. In the discharge of its duties, it may receive complaints related to the sphere of this

Protocol in accordance with the provisions of Articles 259 et seq. of the Criminal Procedure Act.

II.4 Subject to the assessment of the General Council of the Judiciary, the Contact Point could be used to communicate court rulings ordering precautionary measures in matters concerning hate speech crimes to hosting service providers where appropriate and without prejudice to the powers that the judicial authority issuing the ruling already has with regard to notifications.

II.5 The Point of Contact shall use a single dedicated *ad hoc* email address to communicate with hosting service providers. Every hosting service provider must inform the Point of Contact of the specific mechanism required to send it 'notices when acting as a competent authority' by means of that address.

II.6 The removal and conservation of content and its transfer to the competent authorities by hosting service providers shall be governed by applicable regulations and legislation.

II.7 A form agreed to by hosting service providers and the Point of Contact shall be drawn up for notices from the competent authority. This form shall be sent electronically and must contain the following sections:

II.7.A. Preliminary content classification: i.e. whether it is considered illegal content or may be classified as illegal hate speech according to the definition set out in section I.1.

II.7.B. Reliable identification of the Point of Contact as the sender.

II.7.C. The form shall contain the option to request that the content in question be blocked and/or removed within a reasonable period of time. The content described in the notice must be accurately identified and, if technically possible, by means of a uniform resource locator (URL). In order to expedite hosting service providers' decisions, notices must be adequately substantiated so that the hosting service provider in question is able to make an informed and responsible decision. The hosting service provider's decision to block and/or remove content may also be based on notices received from the content provider, where relevant.

II.7.D. To ensure public order and public safety in cases where the hosting service provider decides to remove or block access to content (especially the prevention, investigation, detection and prosecution of serious crimes posing a danger to life or personal safety), the provider may be required to do so confidentially and therefore not inform the content provider of such removal or blocking, or of its reasons, or the possibility of challenging that decision. To that end, the Point of Contact shall indicate in its request to the hosting service provider the period during which it requests such confidentiality, which must be reasonable and proportionate in light of the specific circumstances of the case in question, and under no circumstances exceed 90 days, extendible by means of a new request if such circumstances persist, up to a maximum of an additional 90 days.

III. NOTICES FROM 'TRUSTED FLAGGERS'

III.1 Efforts shall be made so that hosting service providers preferentially process notices from trusted flaggers over those received from ordinary individuals.

III.2 A form will be drawn up for notices from trusted flaggers which shall be sent electronically and must contain the following sections:

III.2.A. Preliminary content classification: an explanation shall be provided of why the content is considered *illegal hate speech* based on the definition set out in section I.1.

III.2.B. Reliable identification of the trusted flagger as the sender.

III.2.C. If the assessment indicates that there is at least an indication that the content constitutes *illegal hate speech*, the form shall offer the options to request blocking, removing or restricting access to the content within a reasonable period of time. The content described in the notice must be accurately identified and, if technically possible, by means of a uniform resource locator (URL). In order to expedite hosting service providers' decisions, notices must be adequately substantiated so that the hosting service provider in question is able to make an informed and diligent decision.

IV. ACCREDITATION AND TRAINING OF 'TRUSTED FLAGGERS'

IV.1 Agreement is reached regarding the definition of 'trusted flagger' laid down in the Recommendation in order to clarify that trustworthiness refers to the fact that flaggers have been accredited by the hosting service provider, 'responsible' means that their activity focuses on issues closely related to combating intolerance and/or discrimination and 'competence' means that they are experienced and have achieved verifiable results in this field.

IV.2 An accreditation procedure shall be established according to which:

IV.2.A. 'Selection criteria' applicable to trusted flaggers shall be freely established by each hosting service provider in accordance with its own policies. Service providers shall be encouraged to publish these criteria on their websites in the form of a clear list of conditions that must be met in order to be selected as a trusted flagger.

IV.2.B. In accordance with the Recommendation, candidates that comply with the policies of each hosting service provider and are therefore selected, shall not automatically be considered trusted flaggers. Once 'trusted flaggers' are chosen, they must meet the 'accreditation criteria' and appear before the 'Trusted Flagger Accreditation Committee' to which they will submit their applications.

IV.2.C. The Trusted Flagger Accreditation Committee, established in compliance with the provisions of the Recommendation, shall be composed of a representative from the Ministry of the Interior —National Office for the Fight against Hate Crime— (who will chair the committee), a representative from the Ministry of Inclusion, Social Security and Migration —Spanish Observatory on Racism and Xenophobia—, a rep-

resentative in Spain from Twitter, YouTube, Facebook and Microsoft, and a representative from a civil society association who is already a trusted flagger for the aforementioned hosting service providers which will rotate annually.

IV.2.D The Accreditation Committee shall evaluate each selected trusted flagger and verify whether the person meets the accreditation criteria. Once this latter verification is approved, the specific hosting service provider that selected this notice provider may consider the latter as a trusted flagger for the purposes of the Protocol.

IV.2.E. As from the time of their first appointment, the Accreditation Committee shall review, on a biannual basis, whether trusted flaggers continue to meet accreditation criteria and they shall lose their status as trusted flaggers if it is found that they do not. Notwithstanding the foregoing, should the trusted flagger fail to pass the status review established by a given hosting service provider in accordance with the procedures and periods laid down in its policies, that service provider shall inform the Committee of this situation and the latter will then proceed to strip that entity of its trusted flagger status with respect to that particular service provider.

IV.2.F. The current list of trusted flaggers shall remain in force for the implementation of these provisions.

IV.3 The Accreditation Committee shall adhere to the following accreditation criteria:

IV.3.A. Evidence of having been selected by the specific hosting service provider, which implies evidence of having passed the training courses related to the operation of the social media in question and other issues required under its internal policy.

IV.3.B. Evidence of having focused their activity over the last three years on issues related to combating intolerance and/or hate

by submitting documents (organization reports, etc.) from which their active participation in a hate counter-narrative (both *online* and *offline*) and efforts to accompany and defend individual victims or groups victimized by hate crimes and hate speech can be inferred.

IV.3.C. Evidence that it is a legal entity with offices in Spain.

IV.3.D. Evidence of having passed a specific training course approved by the Accreditation Committee related to combating online hate speech in accordance with the provisions of the Recommendation.

IV.3.E. When applicable, evidence of other elements not required to achieve trusted flagger status but which will support the candidacy, such as: entities that have already been considered trusted flaggers in the past by that same hosting service provider and/or for other hosting service providers, and that engage in activities to protect groups suffering discrimination or individuals who are traditional victims of hate speech and who do not yet have a trusted flagger recognized as such by a hosting service provider.

IV.3.F. For renewal, evidence of the foregoing points circumscribed to the last two years.

IV.4 Trusted flaggers and hosting service providers shall prepare reports on their activity and forward them to the Accreditation Committee which, in turn, shall regularly inform the Monitoring Committee of the 'Agreement to cooperate on an institutional level in combating racism, xenophobia, LGBTIphobia and other forms of intolerance'.

IV.5 Recognizing that trusted flaggers play an important role in carrying out their business activity in accordance with the law, hosting service providers may directly or indirectly support such work by means of advertising credit or by other means. However, this remuneration may never constitute trusted flaggers' main source of funding.

V. IMPLEMENTATION OF REDRESS MECHANISMS

V.1 The National Office for the Fight against Hate Crime (provided for under Instruction Seven 1/2018) is responsible for informing citizens of alternative approaches outside the scope of criminal courts to settle disputes in the area of hate speech. This does not preclude any other institution or victim assistance office, in accordance with its particular commitments or attributions, from engaging or committing to engage in these information services.

V.2 Where applicable under their internal policies, hosting service providers shall inform the aforementioned Office of the possibility of such alternative approaches.

V.3 Similarly, trusted flaggers shall inform the Office of their degree of satisfaction with such approaches as applicable.

VI. MONITORING AND EVALUATION

VI.1 This *Protocol shall* be incorporated as an Addendum to the 'Agreement to cooperate on an institutional level in combating racism, xenophobia, LGBTIphobia and other forms of intolerance' of 19 September 2018 as provided under Clause Two regarding "collaboration in conducting activities of joint interest".

VI.2 The Monitoring Committee of the *Interinstitutional Agreement* shall monitor the application of the *Protocol* by establishing a collaboration mechanism in which its signatories participate. The annual Chair and the Secretariat of the *Interinstitutional Agreement* Monitoring Committee shall prepare activity reports in relation to the implementation of the *Protocol* which will be sent periodically to the Monitoring Committee.

The President of the Supreme Court and of the General Council of the Judiciary *By proxy. Proxy signature protocol signed on 17 July 2020.* José Antonio Ballestero Pascual	**The State Prosecutor-General** Dolores Delgado García
Secretary of State for Justice Pablo Zapatero Miguel	**Secretary of State for Security** Rafael Pérez Ruiz
Secretary of State for Education Alejandro Tiana Ferrer	**Secretary of State for Sport** Irene Lozano Domingo
Secretary of State for Social Rights Ignacio Alvarez Peralta	**Secretary of State for Equality and against Gender Violence** Noelia Vera Ruiz-Herrera
Secretary of State for Migration Hana Jalloul Muro	**Director of the Centre for Legal Studies** Maria Abigail Fernández González
Spanish Digital Economy Association Carina Szpilka	**State Council of the Roma People** Ignacio Alvarez Peralta
Social Forum for Immigrant Integration Cristina Blanco Fernández de Valderrama	**State Federation of Lesbians, Gays, Transsexuals and Bisexuals** Eugenia Sangil Sánchez
Council for Victims of Hate and Discrimination Crimes Montserrat Moreno Lanza	**Platform of Children's Organisations** Carles López Pico
Third Sector Platform Francisca Sauquillo Perez del Arco	

Co-funded by
the European Union