SJLS
SPANISH JOURNAL OF LEGISLATIVE STUDIES
SPANISH JOURNAL OF LEGISLATIVE STUDIES
ISSN 2695-5792

2020. Artículo 6

FREEDOM OF EXPRESSION IN SOCIAL MEDIA AND CRIMINALIZATION OF HATE SPEECH IN SPAIN: EVOLUTION, IMPACT AND EMPIRICAL ANALYSIS OF NORMATIVE COMPLIANCE AND SELF-CENSORSHIP[1]

Fernando Miró-Llinares
Full Professor of Criminal Law
Miguel Hernández University of Elche

Correspondence:
f.miro@crimina.es

Ana B. Gómez-Bellvís[2]
Researcher FPI at CRÍMINA Centre for the study and prevention of crime
Miguel Hernández University of Elche

Correspondence:
ana.gomezb@umh.es

How to cite this paper

Miró-Llinares, Fernando and Gómez-Bellvís, Ana B. (2020): Freedom of expression in social media and criminalization of hate speech in Spain: Evolution, impact and empirical analysis of normative compilance and self-censorship. *Spanish Journal of Legislative Studies*. (1), p 1-42
DOI: https://doi.org/10.21134/sjls.v0i1.1837

[1] Research conducted within the project "Criminology, empirical evidence and criminal policy. On the incorporation of scientific data for decision-making in relation to the criminalization of conduct - *Reference: DER2017-86204-R, financed by the (AEI)/ MCIU and the European Union through the European Regional Development Fund -FEDER- "A way of doing Europe"*. Also, thanks to grants for dynamization activities "Networks of Excellence" project: *Development of a criminological and empirical model of criminal policy -Acronym EmpiriC. Funded by MCIU-AEI (Ref. DER2017-90552-REDT).*

[2] This research has received funding from the grants for pre-doctoral contracts for the training of doctors 2018 (FPI) from the Ministry of Science, Innovation and Universities and co-financing from the European Social Fund, reference number: PRE2018-083939.

Summary

Freedom of expression in Spain and its determinants. I.1. More or less freedom of expression? Evidence regarding the impact of the criminalization and judicialization of hate speech. I.2. The impact on freedom of expression of the privatization of censorship in social networks: little data, more evidence. II. Crimes of expression in Spain and the judicial response to offences on social networks. II.1. The criminalization of expression since the introduction of the 1995 Criminal Code. II.2. The judicial interpretation of crimes of expression since the popularization of social networks. III. Empirical analysis of regulatory compliance and self-censorship in social networks. III.1. Regulatory compliance, self-censorship and banning expressions on the Internet: measuring the impact of the restrictions. III.2. Empirical study. IV. Results. V. Discussion and limitations

Abstract

The popularization of social media as a forum for the expression of ideas and for political debate has increased in recent years the, always present, tension between freedom of expression and public control of offensive or dangerous speech. Many democratic states criminalise hate speech and other types of offensive expression, and now are social Media themselves that can exercise prior censorship over content on the basis of similar rules but much more restrictive and ambiguous. Combining normative analysis with empirical methodologies, this paper analyses, focusing on Spain, the implications that Criminal laws and "community guidelines" have on citizens' decisions to express themselves freely, which is fundamental for the configuration of a real democracy. It is made up of two studies: the first traces the evolution of judicial resolutions in Spain that prosecute possible crimes of terrorism and hate for political expression on social networks, as a result of the popularisation of these since 2014 in concurrence with a broad and ambiguous legislation. The results show a significant growth of state control over expressions that are essentially offensive but that, in most cases, do not represent a clear and present danger and could go beyond the doctrine of the ECHR. In the second part, an empirical study is carried out to analyse how the existence of hate speech crimes and social media rules affects the decision to express political ideas on social networks or to self-censor. The results show that a large part of the sample is self-censoring but that criminal law and content rules, in particular the severity of the law and the certainty that it is applied, do not have a direct effect on the decision to express ideas on the Internet, while the social perception of what others do is decisive. It is discussed how this calls into question the legitimacy of the evident limitation of freedom of expression of some crimes in Spain.

Key words

freedom of speech, censorship 2.0, self-censorship, social networks

I. FREEDOM OF EXPRESSION IN SPAIN AND ITS DETERMINANTS

I.1. More or less freedom of expression? Evidence regarding the impact of the criminalization and judicialization of hate speech

Recently in Spain, and perhaps due to political tensions, it is common to hear positive or negative public declarations regarding the quality of democracy and in particular the state of freedom of expression, which is an indisputable pillar in the formation and shaping of the democratic being[3]. While the popularization of social networks should improve the free expression of opinions and ideas[4], in reality, many of the aforementioned declarations claim that there has been an involution. It is said that freedom of expression in Spain is at risk[5], that this human right is suffering a regression, and even that public persecution of political dissent is behind the criminalization of certain forms of public expression[6]. Some of the claims are not backed by serious and objective arguments, rather, they merely reference a specific situation which has been partially interpreted. However, others are based on well-known comparative indices that are used to measure the degree of freedom of expression in different countries and can provide support for the alleged regression.

In 2019, *The Human Freedom*[7] Index ranks Spain 29th out of 162 countries[8], one position lower than the previous year. The country received an overall score of 8.69 based on various indexes, including "Laws and regulations that influence the media". This particular indicator is rather low compared to the others relating to freedom of expression and information and has evolved from a score of 8.7 in 2011 to 8.0 in 2019. The evaluation carried out by *Reporters Without Borders* in its 2019 report[9] is of more interest. Despite Spain's

[3] See, for in-depth analysis: PRESNO LINERA, M. Á., and TERUEL LOZANO, G. M., *La libertad de expresión en América y Europa*, Editorial Juruá, Lisboa, 2017.

[4] On the one hand, by increasing possible dissemination and, therefore, the potential impact of any statement, and, on the other hand, by democratizing public declarations by adding social networks to traditional communication channels and, thus, overcoming of some of the limitations regarding access and dissemination of media.

[5] For example, see the headlines in some media sources: "Freedom of expression in Spain at risk, warns Amnesty International" (Available at: https://www.elplural.com/leequid/la-libertad-de-expresion-peligra-en-espana-advierte-amnistia-internacional_124709102); "Bad times for freedom of expression in Spain" (Available at: https://www.huffingtonpost.es/2018/02/22/malos-tiempos-para-la-libertad-de-expresion-en-espana_a_23368416/); "The New York Times warns freedom of expression in Spain at risk" (Available at: https://www.publico.es/internacional/tribunal-europeo-derechos-humanos-the-new-york-times-advierte-libertad-expresion-espana-riesgo.html), among many others.

[6] This was claimed by Amnesty International in their 2018 report entitled "Tweet... if you dare" and in which they reported the persecution of political dissidents through police investigations of statements made on Twitter, which were carried out as part of "operation spider" by the Spanish Civil Guard, and sentenced by the Spanish courts of justice. The report is available at: https://www.amnesty.org/download/Documents/EUR4179242018SPANISH.PDF

[7] The so-called World Freedom Index, the predecessor of the Human Freedom Index, is a report co-published by the Cato Institute, the Fraser Institute and the Friedrich Naumann Foundation's Liberal Institute for Freedom. The report was established in 2015 and analyses freedom in 152 countries on the basis of nearly 80 different indicators of personal and economic freedom.

[8] VASQUEZ, I., and PORNIK, T., *The Human Freedom Index 2019. A Global Measurement of Personal, Civil and Economic Freedom*, 2019. Available at: https://www.cato.org/sites/cato.org/files/human-freedom-index-files/2019-human-freedom-index-update-2.pdf. This is not the index in which our country ranks best. In Western Europe, we are only above Italy and France and below 15 other countries.

[9] Available at: https://www.informeanualrsf.es/

rise from 31st to 29th place and a 1.48 point increase[10], the report expressly highlights that it is a problem that "in 2018 there were convictions for crimes such as: glorifying terrorism, insulting the crown and offending 'the feelings of members of a religious confession', which created a climate that is detrimental to freedom of information". Even more significant is the evaluation of Spain in *Freedom in The World*, which is probably the most widely used and well-known index to measure and compare freedom and democracy in different states and territories[11]. In its 2019 report, the index produced by *Freedom House*[12] gives Spain an overall score of 94 points out of 100, which places it in 18th position (along with 8 other countries). It received one point less than in previous years in the category of Civil Liberties in relation to the evolution of freedom of expression, specifically for item D4 which analyses "Are individuals free to express their personal views on political or other sensitive issues without fear of surveillance or punishment?". The reason explicitly given by *Freedom House* is the existence of "a pattern of using a broadly drafted anti-terrorism law and other legal provisions to prosecute individuals for their political expressions"; and adds that "more aggressive enforcement of laws prohibiting the glorification of terrorism has begun to threaten freedom of expression, with dozens of people, including social network users and various artists, convicted in recent years for what often amounts to satire, artistic expression or political commentary".

If we use the above-mentioned indices to conduct an overall evaluation of the evolution of freedom of expression in Spain, there is clearly room for both pessimism and optimism. It seems valid to say that, on the one hand, compared to other countries Spain is essentially a free country in which people can generally express themselves without fear[13]; while, on the other hand, it can also be said that in recent years it seems that freedom of expression has been harmed. What these instruments do offer is an indication of both potential trends in the regression of freedom of expression, as well as a possible explanation for this involution. As we have seen, when justifying their assessment of Spain, they all refer expressly to the legal and judicial context in which crimes of expression are applied. In particular, they refer to criminal regulation and its interpretation by the courts as a negative indicator of a decline in freedom, and they implicitly reflect the impact that the generalization of social networks as an instrument for political criticism may have had on judicialization. In fact, the indices support those

[10] It should be borne in mind that Spain's performance on this index since 2013 is highly positive. In 2013, it was ranked 36th, but had risen to 29th place by 2017, where it remains today.

[11] Some authors have criticized this index for a neoliberal bias that makes it give higher scores to countries with close ties to the United States or similar political institutions (e.g. GIANNONE, D., "Political and Ideological Aspects in the Measurement of Democracy: The Freedom House Case," in *Democratization*, vol. 17, 2010; STEINER, N. D. Comparing Freedom House Democracy Scores to Alternative Indices and Testing for Political Bias: Are US Allies Rated as More Democratic by Freedom House", in *Journal of Comparative Policy Analysis: Research and Practice*, vol. 18, 2016. However, BARNIDGE, M., HUBER, B., DE ZÚÑIGA, H. G., and LIU, J. H, "Social Media as a Sphere for "Risky" Political Expression: A Twenty-Country Multilevel Comparative Analysis", in *The International Journal of Press/Politics*, vol. 23, 2018, point out that while these studies show a relatively consistent neoliberal bias in the pre-1989 measure, they also show that Freedom House's scores align more closely with other indices of democratic performance after that time).

[12] Available at: https://freedomhouse.org/report-types/freedom-world

[13] The data from the Democracy Index, which places Spain not in the top 10 but in the top 30, and as one of the states that falls into the category of "*fully free*", would also be along these lines. Available at: https://infographics.economist.com/2020/democracy-index-2019/index.html

academics who first warned against increasing criminalization of so-called hate speech and judicialization of offensive expressions in social networks, because of the potential impact this could have on exercising freedom of expression. This was even done without clear knowledge of the decisive role that social networks could play as a means to both exercise freedom of expression and potentially restrict it[14].

However, these indications do not serve to confirm a retrocession of freedom of expression in Spain. Neither do they allow this negative evolution to be aetiologically attributed to the criminalization and judicialization of crimes of expression in the era of social networks. Nonetheless, we can use empirical methodologies to address these ambitious research topics by reducing the research objective to something more modest: on the one hand, to identify tendencies regarding the courts and the crimes provided for in the Criminal Code that are related to the criminalization of political expression through social networks; and, on the other hand, to analyse citizens' perception of freedom of expression and the impact that the criminal repression of so-called hate speech might have on exercising free expression in social networks, where censorship is no longer only a matter for public authorities but also for private entities.

I.2. The impact on freedom of expression of the privatization of censorship in social networks: little data, more evidence.

Analysis of the overall level of freedom of expression includes the potential restrictive effects of formalized state control, such as criminal law or its effective judicial application. These effects may be examined either in general or more specifically in relation to the media, which are essential institutions in the dissemination of information and in the shaping of political pluralism. For a long time, the control exercised over freedom of expression by certain Internet platforms has been ignored, and it is obvious that doing so invalidates any analysis.

Today, the public expression of ideas, including political declarations, is mainly carried out on the Internet. Adopting mechanisms to self-regulate content means social networks also exercise, through their "content policies", control over the free circulation of ideas that, in practice, may even be more restrictive than state control[15].

Although social networks first appeared in the early 2000s, it was not until years later that they began to adopt standards regarding the content allowed on their platforms[16]. At first, these corporations considered themselves to be software companies and were not particularly concerned with the content of their users' statements or

[14] MIRÓ LLINARES, F., "La criminalización de conductas "ofensivas". A propósito del debate anglosajón sobre los "límites morales" del Derecho penal", in *Revista Electrónica de Ciencia Penal y Criminología*, RECPC 17-23, 2015; MIRÓ LLINARES, F., "Taxonomía de la comunicación violenta y el discurso del odio en Internet", in *Revista de Internet, Derecho y Política*, núm. 22, 2016; MIRÓ LLINARES, F. (ED.), *Cometer delitos en 140 caracteres. El Derecho penal ante el odio y la radicalización en Internet*, Marcial Pons, Madrid, 2017.

[15] For in-depth analysis, see: GILLESPIE, T., *Custodians of the Internet. Platforms, content moderation and the hidden decisions that shape social media*, Yale University Press, New Haven & London, 2018; Teruel LOZANO, G. M., "Libertades comunicativas y censura en el entorno tecnológico global", in *Revista de la Escuela Jacobea de Posgrado*, no. 12, 2017.

[16] Moving from a philosophy of "standards" that are identified as the values of the social network itself to one of rules, where they begin to establish what cannot be expressed in the social network, and filter and moderate content (KLONKICK, K., "The New Governors: The People, Rules and Processes Governing Online Speech", in *Harvard Law Review*, vol. 131, 2017).

messages, nor did they take responsibility for it. However, the growth of social networks and their internationalization made it clear that content moderation was necessary. This brought new problems for these networks, such as user dissatisfaction with unjustified withdrawals of content[17], or the use of social networks in different countries with diverse cultures,[18] which meant some governments might block content that was offensive to certain actors within their culture. In addition to these difficulties, there have been various scandals that[19] could undermine the reputation of social networks, as well as demands for public authorities to control possible illegal content produced on social networks by making the networks themselves responsible for the content[20]. This has facilitated the adoption of truly restrictive usage rules, which exercise far greater control (in terms of scope) over what is expressed on their platforms than that exercised by states when they limit freedom of expression. This may either be because public authorities have forced them to establish certain limits on freedom of expression in their respective spaces[21], or to collaborate with governments in the pursuit of radical material, or because it is necessary to maintain their reputation with users and provide a space for communication that is as friendly as possible[23]. Whatever the case, by establishing usage policies, in reality they are exercising truly restrictive control over users' free expression. This may constitute a danger for the plurality of ideas that should be present in any forum for public debate, as ideas may be suppressed through measures such as the removal of messages, or blocking or closure of accounts. The impact of these measu-

[17] GILLESPIE, T., *Custodians of the Internet*...., ob. Cit.

[18] For example, Thailand announced that it would block citizens' access to YouTube if offensive videos against its king were not removed. The type of offenses was not subject to moderation based on the rules established by the platform, but in Thailand, insulting the king is considered a crime (REUTERS. Thailand blocks YouTube for clip mocking king. https://www.reuters.com/article/us-thailand-youtube/thailand-blocks-youtube-for-clip-mocking-king-idUSBKK17066320070404)

[19] On the other hand, Twitter, which had not specified any kind of rules, went from hero to Internet villain, especially because of the GamerGate controversy in 2014, which eventually led to the platform establishing a set of rules and public policies in 2015. (MOTHERBOARD. The History of Twitter's Rules. https://www.vice.com/en_us/article/z43xw3/the-history-of-twitters-rules).

[20] No longer just a conduit for content, the European Commission, together with Facebook, Twitter, YouTube and Microsoft signed a series of commitments in 2016 to combat the spread of illegal hate speech online in Europe, including the removal of content reported as illegal within 24 hours (EUROPEAN COMMISSION. European Commission and IT Companies announce Code of Conduct on illegal online hate speech. https://ec.europa.eu/commission/presscorner/detail/en/IP_16_1937).

[21] In accordance with *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, Tackling Illegal Content Online. Towards an enhanced responsibility of online platforms* (COM 2017, 555 final): "Online platforms should, in light of their central role and capabilities and their associated responsibilities, adopt effective proactive measures to detect and remove illegal content online and not only limit themselves to reacting to notices which they receive. Moreover, for certain categories of illegal content, it may not be possible to fully achieve the aim of reducing the risk of serious harm without platforms taking such proactive measures. The commission considers that taking such voluntary, proactive measures does not automatically lead to the online platform losing the benefit of the liability exemption provided for in Article 14 of the E-commerce Directive" (Available at: https://ec.europa.eu/transparency/regdoc/rep/1/2017/EN/COM-2017-555-F1-EN-MAIN-PART-1.PDF). On detection, see MIRÓ LLINARES, F., "La detección de discurso radical en Internet. Aproximación, encuadre y propuesta de mejora de los análisis de Big Data desde un enfoque de Smart Data criminológico", in ALONSO RIMO, A., CUERDA ARNAU, M. L., y FERNÁNDEZ HERNÁNDEZ, A. (DIRS.), *Terrorismo, sistema penal y derechos fundamentales*", in Tirant lo Blanch, Valencia, 2018.

[22] BOIX PALOP, A., "La construcción de los límites a la libertad de expresión en las redes sociales", in *Revista de Estudios Políticos*, No. 173, 2016.

[23] BALKIN, J. M., "Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation", in *University of California, Davis Law Review*, vol. 51, 2017.

res could be enormous given that the control is prior and that social networks play a leading role in the communication of political debate on a global scale.

The privatization of prior censorship of certain expressions that may be offensive and that, once public, may also be punished by the state has given rise to an intense debate on the scope and limits of private restrictions on the free expression of ideas[24]. This particularly relevant given that these ideas are expressed in a space that is formally private but in practice is public. The academic debate covers multiple topics, including legal analysis of the mechanisms used to filter the messages[25], what type of expressions can be restricted and to what extent, and the crucial question of the democratic legitimacy of these usage rules or policies[26]. Although this article does not aim to resolve these questions, two fundamental premises give meaning to the present study.

The first is based on the recognition that freedom of expression can be limited, including on the Internet, and that rules can be dictated by private entities as well as states[27]. However, any potential limitation must be done from a position that accepts the important role that the expression of ideas plays in the configuration of political pluralism and, therefore, in democracy itself. Given the role social media plays in the expression of political ideas, private entities not only have to be very careful not to harm fundamental rights, but also have to provide themselves with the democratic legitimacy to limit freedom of expression by, for example, adopting transparent decisions based on the maximum possible consensus, and defining their principles and limits in advance.

The second reflection, which actually derives from the previous one, directly concerns the need to restrict the content that can be censored as much as possible, especially taking into account the *chilling effect* that is supposedly associated with any ban on freedom of expression. With reference to the chilling effect, academics who have reflected on the prohibition of expression claim that, given the difficulty to distinguish what is prohibited from what is permitted, any prohibition will always carry the risk that people will self-censor beyond what is established by the norm. If this is the case for public bans, it is likely to also be the case for those implemented by social networks. In this sense, in order to ensure that a certain statement is not restricted or prohibited, it is perfectly possible for citizens to decide not to express their opinions even though they would be protected by their right to freedom of expression.

As we have said, these two reflections form the basis of the empirical analysis to be conducted in the present paper, because they are both normative statements that, nevertheless, are based on factual assumptions that have not been sufficiently analysed by previous literature, much less empirically. On the one hand, we do not know the real extent to which freedom of expression is restricted by the media. On the other hand, we do

[24] STJERNFELT, F., & METTE LAURITZEN, A., *Your Post has been removed. Tech Giants and Freedom of Speech*, Springer Open, 2020; GILLESPIE, T., Custodians..., ob. Cit.

[25] Sobre estas dudas, véase TERUEL LOZANO, G. M., "Libertades comunicativas...", ob. cit. Asimismo, tal y como exponen ZULETA, L., & BURKAL, R., *Hate Speech in the Public Online debate*, The Danish Institute for Human Rights, Denmark, 2017, p. 27, para el caso de Facebook: "Facebook's editing of posts and comments has repeatedly led to intense debate concerning the way in which the editing is carried out in practice. Furthermore, Facebook has been criticised for its non-transparent editing practice, both with respect to statistics available, resources used for editing and translation of guidelines into specific editing practices"

[26] TERUEL LOZANO, G. M., "Libertades comunicativas...", ob. cit.

[27] Veasse ssobre todo esto, en profundidad, PRESNO LINERA, M. Á., y TERUEL LOZANO, G. M., *La libertad de expresión...*, ob. cit.

not know the real impact of the aforementioned *chilling effect* and how it actually functions.

With regard to the first, we must recognize that it is difficult to obtain a full picture of the restrictions on expression in social networks. Despite the fact these platforms have not always been particularly transparent in this respect, in recent years they have published periodic reports that provide some macro data on the main content against which they have acted. However, they do not allow us to examine the extent to which censored specific messages. Twitter and Facebook, which are two of the most relevant networks, both show significant content publication restriction. It should be noted that in the last Twitter report for 2019[28], more than 15 million accounts had been reported, of which only 7,000 had been reported by government entities. This figure may be indicative of the unequal distribution of public and private control over the messages sent and published on these types of platforms. On the other hand, more than five million accounts were reported for infringing the "hate" policy, a policy that is much broader than what is established in our criminal law[29]. Facebook is perhaps the social network that has the most defined content policy and that has the most developed regulation system. This may be either because of it is aware of the importance of regulation for its survival as "the Social Network", or because of its involvement in controversies such as the "the napalm girl" photo or the[30] "Cambridge Analytica" case[31].

In this sense, Facebook offers a little more information in the quarterly reports it has been publishing since 2017, which when taken together with its policy of allowing appeals against decisions to withdraw content and the implementation of an independent "oversight board" to review difficult decisions regarding content policy, shows its concern about these issues. With respect to hate content, data for the third quarter of 2019 indicates that 7 million pieces were removed, nearly double the content removed in the previous quarter for the same reason. Content removed before being reported by users is above 75%, compared to 25% detection in 2017. In terms of "error correction", of the 1.5 million appeals, only 170,000 pieces of content were returned[32]. With regard to terrorist propaganda, slightly more than 5 million items were withdrawn in the last quarter for which we have data, with an automatic detection rate of 99%, there were 134,000 appeals and more than 200,000 items published again, mostly without an appeal. The content withdrawn before being reported by users was above 75%[33]. On the other hand, action was taken against 3 million pieces of content for violating the rules regarding bullying and harassment. Only 16% were automatically detected. More than 700,000 appeals were filed and 100,000 pieces of content against which action had been taken were restored, mostly after an appeal[34]. Perhaps,

[28] Available at: https://transparency.twitter.com/en/twitter-rules-enforcement.html#twitter-rules-enforcement-jan-jun-2019

[29] In addition, more than four million were reported for abusive and harassing behaviour, two million for threats, and nearly two million for messages containing what the social network considers "sensitive media", that is, content that describes or shows particularly sensitive images. According to the same report, action was taken against 500,000 accounts for violating rules regarding hate, against 400,000 accounts for abusive behaviour and 56,000 accounts were reported for threats. See the same at: https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy.

[30] See in this respect the in-depth analysis of GILLESPIE, T.: *Custodians...*, ob. Cit., p. 1 ff.

[31] On the case and in general about the content policies at Facebook and other social networks see STJERNFELT, F., & METTE LAURITZEN, A., YOUR *Post has been removed...*, ob.

[32] See the report at: https://transparency.facebook.com/community-standards-enforcement#hate-speech

[33] Available at: https://transparency.facebook.com/community-standards-enforcement#terrorist-propaganda

[34] Consult at: https://transparency.facebook.com/community-standards-enforcement#bullying-and-harassment

from these last figures, what is noteworthy is that it seems Facebook has certain limitations when it comes to detecting this content automatically in relation to other types of content. This may possibly be a result of the investment made to detect the statements that have caused greatest concern amongst legislators and the international community, specifically statements with radical content that public authorities have expressly requested private platforms to combat.

All this data gives us a measure of the amount of content that is removed. They are a clear indication of the enormous potential impact that decisions to restrict content on social networks can have on the free expression of ideas. But we still need to know a lot to be able to really determine the impact. Thorough analysis of the content that is removed is missing. And even if this were possible, we would still have to analyse the impact these restrictions actually have on the individual decision of citizens to freely express their ideas. And that is the other purpose of this study: to determine, as has been done regarding criminal law, whether and how the prohibitions derived from the application of content policy by social networks affect citizens' decision to express ideas, especially political ideas. In short, it is a question of getting a little closer to understanding the impact that this double normative standard has on the free expression of ideas in order to know, as the 2017 Democracy Index rightly pointed out, if the birth of the Internet has led to a "golden age" or a "golden cage" for freedom of expression.[35]

II. CRIMES OF EXPRESSION IN SPAIN AND THE JUDICIAL RESPONSE TO OFFENCES ON SOCIAL NETWORKS

II.1. The criminalization of expression since the introduction of the 1995 Criminal Code

Within the category of crimes of expression[36], which is made up of crimes whose essential illicit act is the mere verbal or written expression of[37] communicative content, there are many criminal acts, such as, insults and slander, offending religious sentiments, hate speech, or the crimes related to the glorification of terrorism and humiliation of its victims. And not all of them have undergone legislative changes. Neither insults and slander, nor offences against religious sentiments have undergone substantial changes since 1995. This is perhaps because they are offences with strong roots in our Criminal Code and which have been more clearly defined by our legal doctrine and jurisprudence than other crimes of expres-

[35] Available at: http://www.eiu.com/Handlers/WhitepaperHandler.ashx?fi=Democracy_Index_2017.pdf&mode=wp&campaignid=DemocracyIndex2017

[36] This denomination, crimes of expression, has been usually used by the legal doctrine to refer to the crimes of slander and libel (CARMONA SALGADO, C. , "Los delitos de expresión ante la reforma del proyecto de ley orgánica de Código Penal de 1992", in AA. VV, *Política Criminal y Reforma Penal. Homenaje a la memoria del Prof. Dr. D. Juan del Rosal*, Editorial Revista de Derecho Privado, Madrid, 1993); also more broadly to include hate crimes or glorification of terrorism (MIRÓ LLINARES, F. (DIR.), *Cometer delitos en 140 caracteres...*, ob. cit.), and also used by the Criminal Policy Study Group, which includes the following in its proposal for reform of expression offences: Insult and slander of all types; crimes that provoke hate crimes and terrorism; and, crimes against religious sentiments (GRUPO DE ESTUDIOS DE POLÍTICA CRIMINAL, *Una propuesta alternativa de regulación de los delitos de expresión*, Tirant lo Blanch, Valencia, 2019), although with regard to the latter crimes, hate, glorification and offending religious sentiments, have also been called crimes of opinion (ALASTUEY DOBÓN, C., " Discurso del odio y negacionismo en la reforma del Código Penal de 2015", in *Electronic Journal of Criminal Science and Criminology*, RECPC 18-14, 201)

[37] And that also includes symbolism or, thanks to social networks, any content that comes to express that which the rule seeks to prevent being expressed.

sion. The same cannot be said, however, for crime of hate speech in article 510 or the crimes of glorification of terrorism and humiliation of its victims in article 578, perhaps because these are offences that are closely linked to certain social and political contexts.

Since its introduction in the legal system, the hate speech offence provided for in Article 510 has been one of the most controversial crimes in legal literature[38], and was classified as a crime of opinion from the very beginning[39]. Although the most immediate precedent for this precept is found in articles 165 ter and 137 bis b)[40] introduced by Act no. 4/1995, of 11 May, [41] which criminalized glorification of hatred and genocide, it was the 1995 Criminal Code that introduced article 510.1. This broadened the scope of the punishment to include direct and indirect[42] provocation or incitement. Likewise, provocation of genocide also underwent a profound change in that mere denial or justification of genocide became a crime in the well-known article 607.2, which was later partially annulled by the Constitutional Court in the famous STC 235/2007 (Librería Europa case)[43]. Despite all the interpretations in the legal doctrine, as well as the criticism and attempts at a restrictive interpretation of these types of crimes[44], the 2015 reform of the Criminal Code introduced a new article 510[45]. This increased the scope of punishment to unforeseeable levels[46], as it criminalized preparatory acts and provided for the same punishments, thereby linking criminalization to the alleged creation of a particular climate. Expressions that may cause hatred or hostility were also criminalized (despite the difficulty to conceptualize both terms and for criminal law to effectively criminalise feelings), the corresponding punishments increased, and aggrava-

[38] See MIRÓ LLINARES, F. , "Derecho Penal y 140 caracteres. Hacia una exegesis restrictiva de los delitos de expresión", in Miró Llinares, F. (Dir.), *Cometer delitos en 140 caracteres...*, ob. cit.

[39] ALASTUEY DOBÓN, C. , "Discourse of Hate...", ob. cit.

[40] which introduced the crime of advocacy of hatred in article 165b and of genocide in article 137a(b)

[41] The reasons for its introduction respond to the "proliferation in different countries of Europe of episodes of racist and anti-Semitic violence that are perpetrated under the flags and symbols of Nazi ideology" but also to the international obligations assumed by Spain (see in depth AGUILAR GARCÍA, M. A. (DIR.), GÓMEZ MARTÍN, V., MARQUINA BERTRÁN, M., DE ROSA PALACIO, M., TAMARIT, J. M., and AGUILAR GARCÍA, M. A, *Manual práctico para la investigación y enjuiciamiento de los delitos de odio y discriminación*, Centre d'Estudis Jurídics i Formació Especialitzada, 2015) and to the Constitutional Court's own jurisprudence, which, at the beginning of the 1990s, already leaned towards restricting the right to freedom of expression of certain denialist, racist and discriminatory speeches (LAURENZO COPELLO, P., "La discriminación en el Código Penal de 1995", in *Estudios Penales y Criminológicos*, No. 19, 1996). Especially, with respect to the famous STC 214/1991, of November 11 (Violeta Friedman Case), and which is expressly cited in the Explanatory Memorandum of O.L. 4/1995, amending the Criminal Code). In the same sense, Gómez Martín, V., "Discurso del odio y principio del hecho" in MIR PUIG, S. , and CORCOY BIDASOLO, M., *Protección penal de la libertad de expresión e información. Una interpretación constitucional*, Tirant lo Blanch, Valencia, 2012.

[42] By eliminating the term "direct" required by the previous hate crime.

[43] On this, see ROLLNERT LIERN, G., "Revisionismo histórico y racismo en la jurisprudencia constitucional: los límites de la libertad de expresión (a propósito de la STC 235/2007)", in *Revista de Derecho Político*, No. 17, 2008.

[44] For in-depth analysis, see: MIRÓ LLINARES, F., "Derecho Penal y 140 caracteres", ob. cit.

[45] On the exegetical interpretation of this type of criminal *law*, see RODRÍGUEZ FERRÁNDEZ, S. , "El ámbito de aplicación del actual art. 510 CP en retrospectiva y en prospectiva tras la reforma penal de 2015", in *Revista de Derecho Penal y Criminología*, no. 12, 2014. See also PORTILLA CONTRERAS, G. El retorno de la censura y la caza de brujas de anarquistas", in MIRÓ LLINARES, F. (DIR.), *Cometer delitos en 140 caracteres...*, ob. cit.

[46] As TERUEL LOZANO points out, this goes far beyond what the 2008 Framework Decision established and demanded from Member States and even beyond what the TC established in STC 235/2007 (TERUEL LOZANO, G. , "La libertad de expresión frente a los delitos de negcionismo y de provocación al odio y a la violencia: sombras sin luces en la reforma del Código Penal", in *Indret*, no. 4, 2015).

ting circumstances introduced for when they are committed through ICTs[47]. In addition, there were many other issues that are difficult to fit into a liberal legal system that respects freedom of expression, as analysed in depth elsewhere,[48] and whose constitutionality has been questioned by some authors[49].

The crime of glorifying terrorism and humiliating its victims has also been the subject of in-depth analysis in recent years by legal literature[50], which has been particularly concerned about the proliferation of convictions for offensive messages on social networks[51]. This offence has been interpreted as a "different species"[52] because it involves what the courts have described as a ban on hatred[53]. This is essentially the kind of behaviour provided for in article 578 of the Criminal Code. This article punishes both the glorification of acts of terrorism or their perpetrators and the humiliation of their victims and, although it is placed alongside the offences of terrorism in the Criminal Code, as legal literature has repeatedly stated, it does not punish pure acts of terrorism but their glorification.

The crime of exaltation *strictu sensu* actually has its antecedents in the Criminal Code of 1973, under the umbrella of apologia, which was regulated in different precepts, including apologia (public defence) of terrorism[54]. Later, the Criminal Code of Spanish democracy was modified when Act No. 7/2000 of 22 December 2000 introduced the offence of glorification of terrorism and humiliation of its victims for the first time. Its introduction was essentially a response to a very specific social context in our country: the nationalist terrorism of the terrorist group ETA, who-

[47] See the interpretation and analysis of the crime in MIRÓ LLINARES, F., "Derecho Penal y 140 caracteres...", ob. cit.

[48] Ibid.

[49] TERUEL LOZANO, G., "La libertad de expresión frente...", ob. cit. This new regulation is also justified by the legislator in the need to transpose Framework Decision 2008/913/JHA, but clearly this criminalization of hate speech goes far beyond this international regulation. In the same vein, RODRÍGUEZ FERRÁNDEZ, S., "El ámbito...", ob. cit. For an analysis of international hate speech regulations, see ROLLNERT, G., "El discurso del odio: una lectura crítica de la regulación internacional", in *Revista Española de Derecho Constitucional*, No. 115, 2019. The same author, on hate in social networks from an international perspective in ROLLNERT, G. , "Redes sociales y discurso del odio: perspectiva internacional", in Revista de *Internet, Derecho y Política*, 2020, in press. Manuscript provided by the author.

[50] See MIRÓ LLINARES, F., " Ofender como acto de terrorismo. A propósito de los casos "César Strawberry" y "Cassandra Vera ", in DE LA CUESTA AGUADO, P. M., RUÍZ RODRÍGUEZ, L. R., ACALE SÁNCHEZ, M., HAVA GARCÍA, E., RODRÍGUEZ MESA, M. J., GONZÁLEZ AGUDELO, G., MEINI MÉNDEZ, I., & RÍOS CORBACHO, J. M. (COORDS.), *Liber amicorum: Estudios jurídicos en homenaje al profesor doctor Juan María Terradillos Basoco*, Tirant lo Blanch, Valencia, 2018.

[51] See in depth MIRÓ LLINARES, F. (DIR.), Cometer delitos en ..., ob. cit. Also, GÓMEZ MARTÍN, V., "Odio en la red. Una revisión crítica de la reciente jurisprudencia sobre Ciberterrorismo y Ciberodio", in Revista de Derecho Penal y Criminología, no. 20, 2018; GALÁN MUÑOZ, A., "El delito de enaltecimiento terrorista. ¿Instrumento de lucha contra el peligroso discurso del odio terrorista o mecanismo represor de repudiables mensajes de raperos, twitteros y titiriteros", in Estudios Penales y Criminológicos, vol. 38, 2018.

[52] LANDA GOROSTIZA, J. M., " Incitación al odio: evolución jurisprudencial (1995-2011) del artículo 510 CP y propuesta de lege lata (A la vez un comentario a la STS 259/2011 -librería Kalki- y a la STC 235/2007)", in *Revista de Derecho Penal y Criminología*, no. 7, 2012.

[53] For all of them, STS 812/2011 of 21 July.

[54] Article 268 (Common provisions on offences relating to terrorism) reads: "Public advocacy, whether oral or written, or by means of the printing press or other dissemination of the offences covered by this title, and of those guilty of such offences, shall be punishable by a minor term of imprisonment".

se support, in the form of particular expressions, could potentially cause offence[55]. Prior to this reform, it had been understood that apologia required direct incitement, which requires provocation in such a way that the conduct constituted a threat to a specific legally protected right or asset. However, when the glorification of terrorism was introduced as a form of apologia or public defence, most of the legal doctrine considered that this new precept included indirect incitement[56]. Thus, the punitive scope was considerably expanded by providing for a crime that, unlike apologia of terrorism found in article 18 of the Criminal Code, made it possible to punish the mere glorification or public justification of terrorist acts, therefore not requiring express and direct encouragement for the commission of a terrorist offence[57]. Subsequently, the reform introduced by Act no. 2/2015, of March 30 modified this precept in such a way that the new article 578 CP increased sentences for the criminalized conducts and added two possible aggravating circumstances. The crime could be considered aggravated and thus the sentence increased to the highest possible level, when, on the one hand, the act is carried out through ICTs, or, on the other hand, it is considered that the actions are sufficient to seriously disturb public peace or create a serious feeling of insecurity or fear in society or part of it.

II.2. The judicial interpretation of crimes of expression since the popularization of social networks

Over the years, we have seen that and as far as hate crimes and the glorification of terrorism are concerned, there has been an expansion of what is punishable. However, such legislative developments do not necessarily have a real impact on the law in practice. In order to know whether the combination of increased criminalized conducts and use of social networks has meant an increase in trials and in convictions and acquittals, as indicated by the aforementioned indices, we have carried out an exploratory study of sentences. A total of 217 sentences located in the Aranzadi database have been analysed: 46 corresponding to the application of article 510; 150 regarding article 578; 14 regarding articles 490.3 and 491.1; and, 6 for article 525[58]. Figure 1 shows the trends

[55] This is justified by the Explanatory Memorandum to Organic Law 7/2000 of 22 September, which argues that: "it is a matter of something as simple as pursuing the exaltation of terrorist methods, which are radically illegitimate from any constitutional perspective, or of the perpetrators of these crimes, as well as the particularly perverse conduct of those who slander or humiliate the victims while increasing the horror for their relatives. All these acts cause perplexity and indignation in society and deserve a clear criminal reproach".

[56] BERNAL DEL CASTILLO, J., "El enalcimiento DEL terrorismo y la humillación a sus víctimas como formas del "discurso DEL odio", in *Revista de Derecho Penal y Criminología*, no. 16, 2016, p. 15,

[57] GALÁN MUÑOZ, A., "El delito de enaltecimiento...", ob. cit. Likewise, this punitive extension was not only confirmed but reaffirmed by the judicial interpretation of the type. On this, see in depth CARBONELL MATEU, J. C. "Crisis del garantismo penal y el papel de los penalistas", in VV.AA., *Estudios jurídico penales y criminológicos en homenaje a Lorenzo Morillas Cueva*, Dykinson, Madrid, 2018, p. 86.

[58] The criteria for inclusion were, on the one hand, with regard to the provisions examined, O.L. 10/1995 of 23 November, and the corresponding articles, from 23 November 1995 to 18 February 2020. With regard to the type of ruling, only the sentences from criminal proceedings have been accepted. In this way, and using these criteria equally in the four crimes, a total of 217 sentences were obtained (after eliminating from the sample those sentences that were duplicates or that, despite the search terms, did not correspond to the crimes examined). Thus, it should be noted that in no case is it being stated that these are all the sentences that exist, since this data is not available through a database such as Aranzadi. Nevertheless, we can take this data as an indicator or trend for judicial decisions, although always with the necessary caution that an exploratory analysis of this type requires.

in the judicial application of these crimes in the last 20 years[59]. First, it should be noted that except for the crime of glorification of terrorism and humiliation of its victims in Article 578 and, to a lesser extent, hate crimes in Article 510, the rest have been applied relatively little. For this reason, we will only analyse these two crimes insofar as the increase that can be observed in the graph allows for a better analysis of the possible impact of social networks on the prosecution of crimes of expression[60].

With regard to the crime of hate speech, from the sample analysed it can be seen that in the 17 years from 1999 to 2016 we only have 15 sentences, while in the three years from 2017 to 2019 there are 31 sentences. Without information on the total number of cases prosecuted, it is not possible to know if this is a general trend, but it seems unlikely that it is not related to the impact of social networks. Thus, although there were some prosecutions for hate speech carried out on the Internet,[61] it is not until 2016-2019 that there

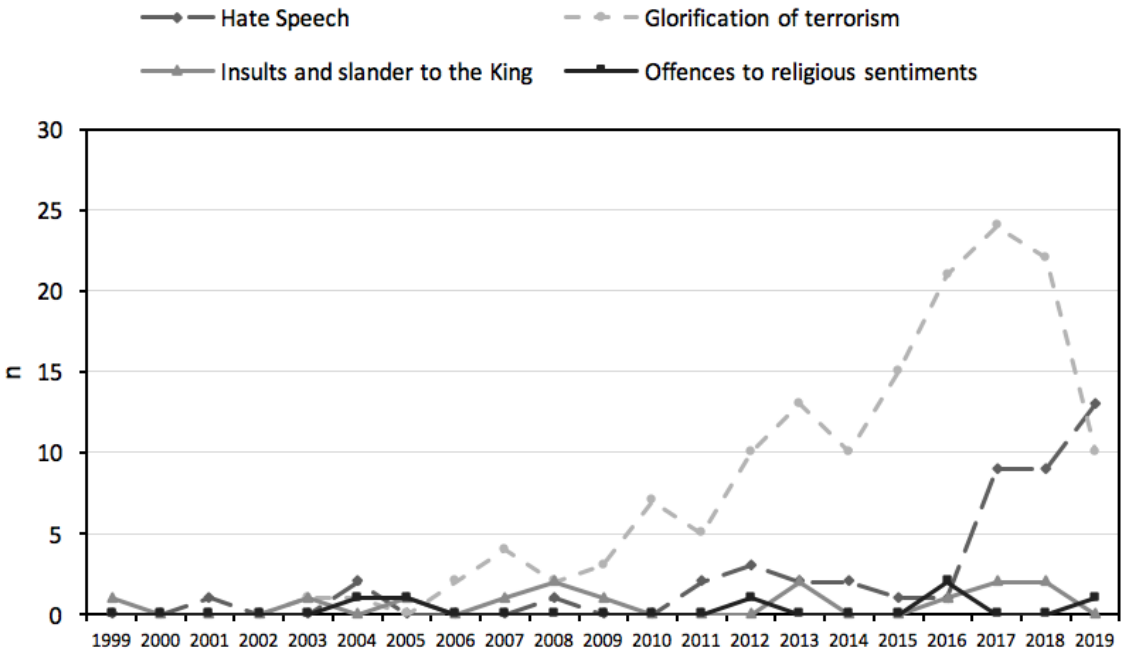

Figure 1. Court decisions by year.

[59] In relation to this graph, it should be noted that the judicial sentences obtained with the inclusion criteria have been considered and where the rulings of the various jurisdictional bodies are included, both Criminal Courts, Provincial Courts, National Court, High Court of Justice and Supreme Court. Therefore, this graph does not show the trend in cases, but rather the trend in judicial rulings, that is, judicial activity in the referenced time frame.

[60] For a detailed analysis of the crime of offences against religious feelings, see RAMOS VÁZQUEZ, J. A., "Muerte y resurrección del delito de escarnio en la jurisprudencia española", in *Revista Electrónica de Ciencia Penal y Criminología*, no. 21, 2019. See also ALCÁCER GUIRAO, R., "Cocinar cristos y quemar coranes. Identidad religiosa y Derecho penal", in MIRÓ LLINARES, F. (DIR.), *Cometer delitos en 140 caracteres...*, ob. cit.

[61] This would be the case of the Sentence from Criminal Court number 2 of Vigo number 22/2012, of January 24th where the individual was sentenced for a crime of justification of genocide for uploading to a web page several photos of themselves with Nazi symbols and making anti-Semitic expressions; the sentence from Criminal Court number 7 of Palma de Mallorca, no. 419/2012, in the case of a video game uploaded to a website and entitled "20 ways to kill a woman", in which the accused was initially convicted and later acquitted by the SAP Balearic Islands, no. 312/2013, of 10 December.

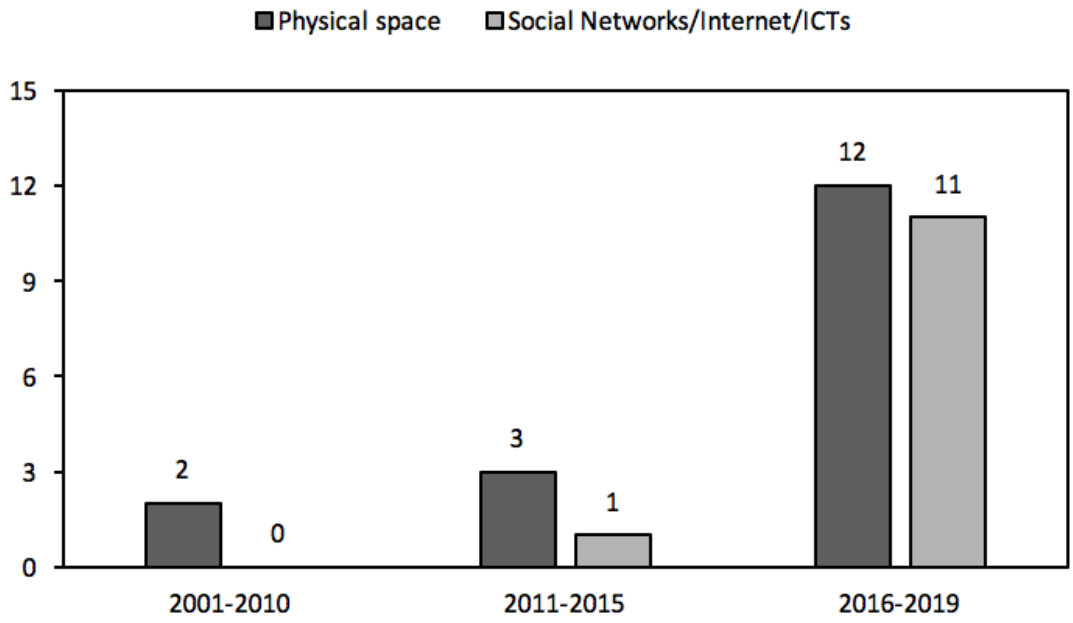

Figure 2. Place where expressions that could be considered hate speech are shared.

is a clear increase in expressions that can be classified as "hate speech" through social networks[62], as shown in Figure 2, which only takes into account the decisions of the Provincial Courts and the National Court[63].

With regard to provocation or incitement to hatred, discrimination or violence, before it was amended and extended by the above-mentioned 2015 reform, the courts had required that the expressions directed against the groups protected[64] by our Criminal Code involved provocation consistent with direct incitement to commit an

[62] For example, the Sentence from Criminal Court No. 1 of Pamplona No. 273/2016, October 11, for uploading a video with anti-Semitic content on Facebook; the SAN No. 2/2017, January 26, for publishing expressions on Twitter such as "53 women murdered by gender violence so far this year, which only seems like a few to me given the number of whores about"; the Sentence of the Court of Instruction No. 8 of Cerdanyola del Vallés, num. 25/2017, de 15 de marzo por realizar expresiones en Twitter de lo que se ha llamado "odio a los catalanes" o "catalanofobia"; la sentencia del Juzgado de lo penal núm. 22 of Barcelona of 11 September 2017 for disseminating through its website opinions and documents that denigrate and humiliate immigrants who profess the Muslim religion, Jews and persons of African descent, defending at all costs the supremacy of the Aryan race over others; SAP Madrid No. 762/2017 for publishing a video on the social network entitled "Sodomy and paedophilia are two branches of the same trunk"; and others.

[63] For those cases in which the accused make expressions that can be classified as both a hate crime and an offence to glorify terrorism.

[64] With regard to protected groups, it is necessary to mention that they are those persons or groups of persons who may be discriminated against for racist, anti-Semitic or other reasons relating to ideology, religion or beliefs, family status, the membership of an ethnic group, race or nation, their national origin, sex, sexual orientation or identity, on grounds of gender, illness or disability. It is important to point this out because in the last available resolutions, the possibility was raised of legally qualifying as a hate crime injurious expressions made on the social network Facebook against a deceased bullfighter, a matter that was not accepted in accordance with the Sentence of the Criminal Court No. 1 of Segovia, No. 419/2019, of 15 November.

offence by words or means, regardless of their actual effectiveness[65]. Furthermore, when the expressions have been disseminated via social networks, the courts have argued that the conduct is more dangerous because of the place where it is expressed,[66] and have even argued the existence of malice[67].

With respect to the crime of glorification of terrorism and humiliation of its victims, if we only consider the National Court sentences, Figure 3 shows that during the first ten years of this crime's existence, the number of sentences is not particularly alarming in quantitative terms, and most of them are related to glorification carried out in physical space[68]. However, the place where the act is committed begins to change in

[65] Ruling of the Criminal Court No. 2 of Logroño No. 133/2004, of 2 April, condemning the distribution of leaflets with phrases such as "we are condemned to live with the garbage of immigrants who will end up destroying us..."; "Moors, South Americans, Eastern countries, Pakistanis, Indians, Africans, etc. All this rabble has more rights than any Riojan", etc. The requirement of direct incitement to appreciate the crime of provocation to discrimination, hatred or violence also in the SAP Barcelona, of 5 March 2008 (Librería Europa case, after the question of unconstitutionality resolved by STC 235/2007) Likewise, SAP Santa Cruz de Tenerife No. 107/2014 of 7 March recalls that "the use of the term provocation in the wording of the first paragraph of article 510 of the punitive text has led to the argument that the requirements of article 18 must be met, except for the requirement that the act to which it is provoked constitutes an offence, since by including provocation to hatred reference is made to a feeling or emotion whose mere existence is not criminal. According to this criterion, it must in any case be a direct incitement to the commission of minimally specific acts which may be preached by discrimination, hatred or violence against the said groups or associations and for the reasons specified in article [...] According to the case law, the following defining elements must be given: (a) The initiative to carry out one or more criminal acts, not just vague and generalised encouragement; (b) the recipient's perception of the words or means of encouragement; (c) the fact that the encouragement is of a persuasive and persuasive nature".

[66] Thus, for example, the judgement of Barcelona Criminal Court No. 22 of 11 September 2017 condemning a person of Nazi ideology for incitement to hatred on the Internet, expressly states that "this court attaches particular importance to the medium used and the context in which the incriminated texts were disseminated in this case and, consequently, to the potential impact. This is not a flyer or a speech, but a web page that has been visited more than 30,000 times in five months. The use of a medium as powerful and widespread as the Internet is a totally suitable and rapid channel for the propagation of ideas with such content to reach a large number of people, regardless of their geographical location, and is likely to stimulate in them a state of opinion of animosity, and sometimes hatred, towards the groups mentioned". Likewise, this amplified publicity offered by the Internet is also taken into account, but in the opposite direction, the SAP Barcelona No. 299/2019 which confirms the sentence for a hate crime in paragraph 510.2 after the reform of the Criminal Code of 2015, but instead of applying the aggravating circumstance for the use of ICTs decides to apply the basic rate because by disseminating the hate messages through a Whatsapp group, the Court considers that 60 people, who were the members of the group, are not quantitatively many people.

[67] Thus, the sentence of the Criminal Court No. 1 of Pamplona No. 273/2016, of October 11, that in view of the defendant's claim that he had no knowledge of what he was uploading on the social network Facebook, it is argued that "the dynamic on Facebook is precisely to share comments, images, opinions or information with third parties, given that it is a social network. The accused alleged that he did not remember to post it, that he did it without realizing it and that it is against violence, but the dynamics of Facebook makes it unlikely that he would post something on his own wall without realizing it, especially since the accused knows how it works, since he had two profiles, differentiating one that was public and one that was private, and the public of the accused that is in the car contains many other videos".

[68] Related to expressions in demonstrations, parliamentary statements, banners, leaflets or popular festivals or expressions made on the occasion of the death and burial of supporters and members of the terrorist group ETA. This is the case, for example, with the: STSJ del País Vasco, 5 September 2003; STSJ del País Vasco, 31 March 2004; SAN No. 31/2006, 26 April; STS No. 585/2007, 20 June; SAN No. 67/2007, 12 November; SAN No. 49/2008, 29 July; STS No. 539/2008, 23 September; SAN No. 539/2008, 23 September; STS No. 585/2007, 23 June; STS No. 585/2007, 23 June; SAN No. 585/2007, 24 November; SAN No. 585/2007, 24 April; SAN No. 585/2007, 24 June; SAN No. 67/2007, 12 November; SAN No. 49/2008, 29 July; STS No. 539/2008, 23 September. 28/2009 of 21 May; STS No. 676/2009 of 5 June; SAN No. 64/2009 of 16 December; SAN No. 1/2010 of 19 January; SAN No. 13/2010 of 2 March; SAN No. 224/"010 of 3 March; SAN No. 54/2010 of 9 December, and many others.

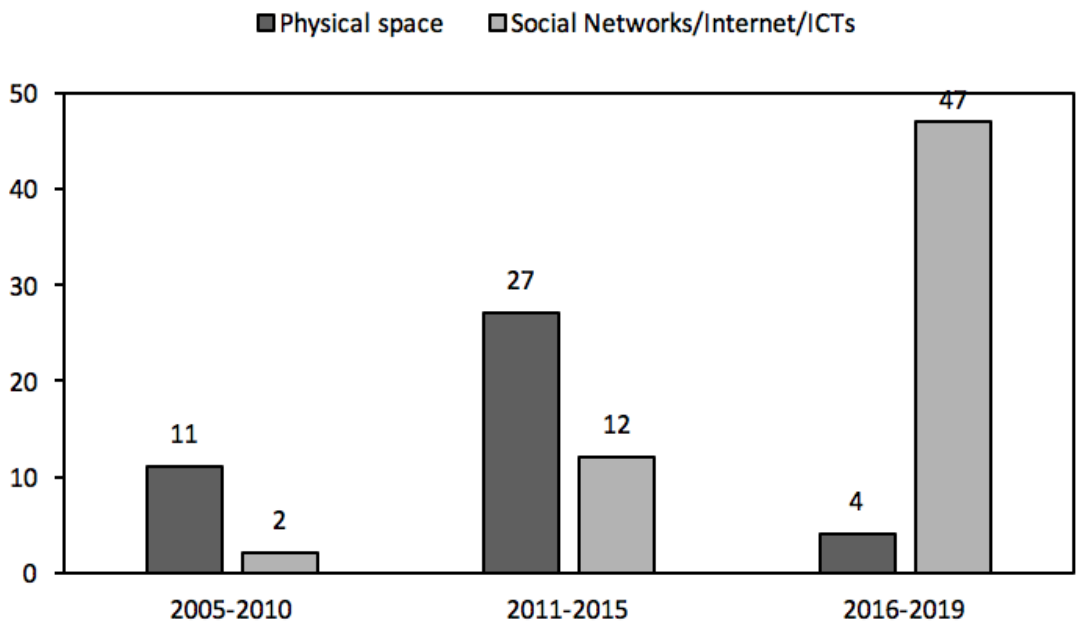

Figure 3. Place where expressions that can be considered crimes of exaltation of terrorism and humiliation of its victims have been published.[72]

2014[69] when the number of sentences and prosecutions for expressions made mainly through social networks such as Youtube, Facebook or Twitter begins to increase, albeit moderately[70]. There are a notably high number of sentences between 2016 and 2020. In this regard, it is particularly noteworthy that between 2005 and 2010, when the terrorist group ETA was active and wreaking havoc on Spanish society, there were only 13 sentences, mostly physical space; while in the period 2016 to 2020, with the terrorist organization happily defunct, there were 51 sentences, 47 of which involved the prosecution of expressions made on social networks. Although it is not possible to categorically state that social networks have increased the commission of crimes, it does seem reasonable to comprehend that what has increased with the appearance of social networks is the potential publicity of what is being expressed, the greater exposure of opinions and, based on the above, the increased control by the investigative bodies[71].

Likewise, without all the sentences or the number that this sample represents of the total,

[69] Although there were some sentences that prosecuted the expressions made through the Internet: SAN No. 62/2006, of 21 November, which corresponds to the acquittal of the members of the music group Sociedad Alkoholika for their songs uploaded to a website; SAN No. 4/"010, of 2 March, which prosecutes expressions made in Internet forums; SAN No. 2/2012, of 17 January, which prosecutes certain comments on the Tuenti social network; and SAN No. 11/2012, of 29 February, which prosecutes messages against a female Euro-parliamentarian on the Internet.

[70] It's the case of: SAN No. 8/2014 of 31 March; SAN No. 24/2014 of 19 May; SAN No. 13/2015 of 20 May; SAN No. 14/2015 of 25 May; SAN No. 37/2015 of 12 June; SAN No. 39/2015 of 14 October; SAN No. 56/2015 of 16 October; SAN No. 32/2015 of 23 November.

[71] This may be what happened with the well-known "operation spider".

[72] For Figures 3 and 4, only the sentences from the National Court have been used to identify trends and not to analyse individual cases. In this sense, it is necessary to take into account that some persons who the National Court has convicted for this crime, have subsequently been acquitted by the Supreme Court, as in the case of Cassandra Vera.

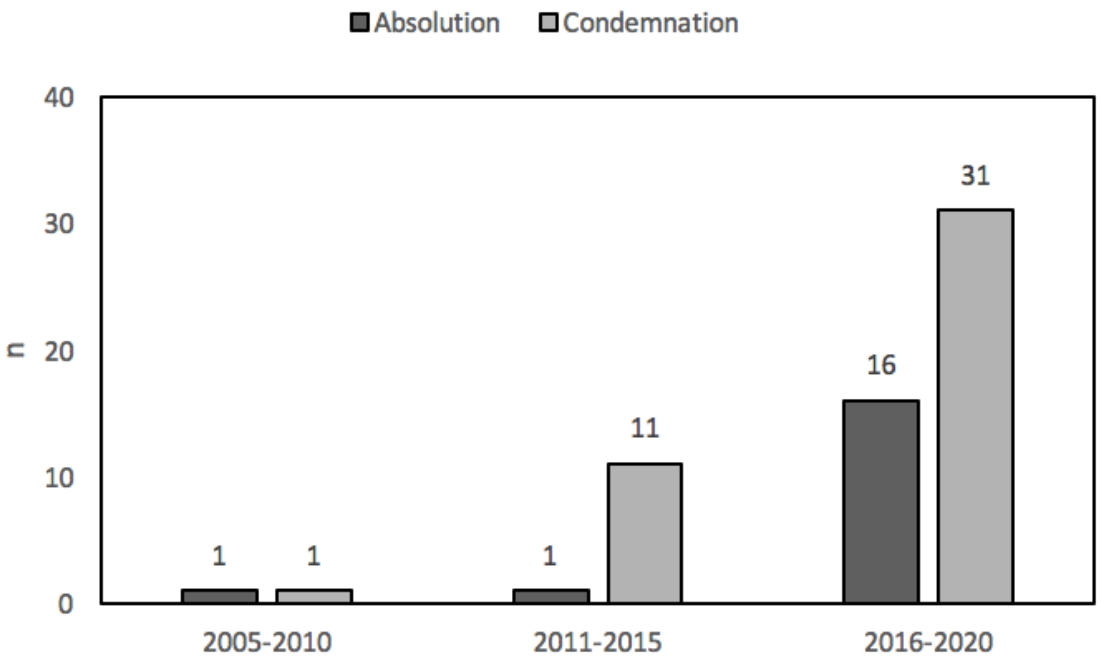

Figure 4. Trend in court sentences for expressions on the Internet and social networks

it is risky to examine whether in addition to the growth in the number of trials there is also a proportional increase in the number of convictions. But it is not risky to say that since 2014 there are many more people convicted of this crime than in the previous period of its existence. If we look only at sentences from the National Court that have prosecuted expressions made on the Internet and social networks, in Figure 4 we can observe that this court tends to convict for this crime[73].

In addition to the elements constitutive of the offence, sentencing normally takes into account the activity carried out on the social network, that is, the specific number of offensive tweets or messages, the total number of tweets or messages sent from the account, the number of followers[74], and how public the user's account is[75]. This activity is used to determine the level of danger presented by the expression[76].

In the case of glorification, on the one hand, the objective elements (actus reus) that constitu-

[73] This is the conviction for expressions such as "Spain is our routine, to beat and burn Spain our routine; the struggle is the only way" (SAN no. 3/2016, 23 February); "Up with those guys that have brought down the capitalist branches in the city of Bilbao. Damn, with news like that it is good to start the week" (SAN no. 4/2016, March 1); referring to the mother of an ETA victim "what can we expect from a bad mother who does business by supporting the impunity of her son's murderers? (SAN No. 25/2017, December 4); or, "SOS ETA; 1, 2, 3, English duck, without moving hands and feet. Irene Villa behind; I need a stamp from Carrero Blanco, the first Spanish astronaut, and I need it yesterday" (SAN no. 3/2018, January 15).

[74] SAN No. 9/2017 of 29 March; SAN No. 4/2018 of 18 January

[75] SAN No. 18/2017 of 21 July

[76] SAN 14/2015, 25 May: "Today social change is linked to technological evolution. The Internet has brought about a revolution in the world of communications and knowledge, with the particularity that it allows backward countries to advance enormously, but this importance of social networks also has an impact on the penal system. Any criminal policy today cannot ignore this technological explosion that allows any message to be disseminated in a few seconds to a multitude of users located in distant countries, thereby obtaining publicity for the messages that would have been unthinkable a few years ago".

te the offence are: (a) words or expressions that glorify or justify in a broader sense than those established in article 18 of the Criminal Code of (b) terrorist conduct or its perpetrators and (c) via any means of public dissemination[77]. On the other hand, there must be a subjective element (*mens rea*) that according to the Courts interpretation does not involve any real intention to provoke violence, but simply intent to knowingly and willingly carrying out the previous objective elements[78]. In the case of "humiliation of the victims of terrorism or their families", in short, it involves insults or slander that are aggravated as a consequence of the person to whom they are directed. Therefore, publicity is not required as an element of the offence, but simply "acts" or the expression of messages that objectively imply a manifest disregard for the victims of terrorism, to such an extent that they can be said to have been humiliated and degraded[79]. The same mens rea is required as in the case of the glorification of terrorism. In this sense, if the elements constitutive of the crime could be identified in the expression or statement, the conviction was practically assured, regardless of whether they were really glorifying in the sense of calling for violence and creating a legally prohibited, albeit abstract, threat to a specific legal right or asset, or if there was no real intention to humiliate or offend certain victims or their relatives, or no knowledge and willingness to create such a risk.

However, it is important to note a change in the jurisprudence regarding the interpretation of glorification in light of the interpretation of Article 578 of the Criminal Code by the Constitutional Court in STC 112/2016, June 20[80]. Even though it is not found in the literal text of the article, in order to limit freedom of expression there now must be a risk, albeit indirect, for people or the rights of a third party derived from the incitement, as required by international law[81] and the ECtHR itself[82]. Likewise, and although it is still too early to be able to say this more restrictive interpretation will be consolidated, we could predict that this will end up being the case if we take into account the recent ruling by the Constitutional Court[83] that annuls the Supreme Court's sentence 4/2017, of 18 January. This had sentenced César Augusto Mon-

[77] STS No. 149/2007 of 26 February 2007, among others.

[78] STS No. 948/2016 of 15 December: "This Court has stated in previous resolutions that the subjective elements of the criminal types are accredited by inference trials can be considered as psychic facts that can be inserted in the factual narrative of the sentence. It is stated that the subjective elements must be deduced from external and objective data contained in the factual account"; SAN No. 5/2018, 18 January.

[79] SAN No. 37/2015 of 12 June.

[80] Which in turn refers to STC 177/2015, of 22 July and develops the three elements that characterize the right to freedom of expression: a) institutional character; b) limitability, and therefore it is admissible for democratic societies to sanction and even try to prevent all forms of expression that propagate, incite, promote or justify hatred based on intolerance; c) proportionality in the limitation of freedom of expression, which means taking into account the risks derived from the use of ius puniendi in the State's response to freedom of expression, whether or not it is excessive, due to the potential disproportionate use of this power and the chilling effect that it may generate.

[81] Especially as regards Directive 2017/541 which in its tenth recital states with regard to public provocation (which includes glorification and justification of terrorism or the dissemination of messages or images, including those relating to the victims of terrorism) "should be criminalised when it means there is a risk terrorist acts may be committed. In each specific case, the examination of whether such a risk has materialised should take into account the specific circumstances of the case, such as the author and the recipient of the message, as well as the context in which the act has been committed. The significance and credibility of the risk should also be considered when applying the provision on public provocation in accordance with national law".

[82] For all, STEDH of 2 October 2008, Leroy v. France.

[83] The sentence can be consulted at the following link: http://www.tribunalconstitucional.es/NotasDePrensaDocumentos/NP_2020_035/2017-2476STC.pdf

taña Lehman, better known as César Strawberry for tweets that were deemed to glorify terrorism as per article 578 of the Criminal Code[84].

The Constitutional Court ruling essentially applies what is set out in the aforementioned STC 112/2016 and gives pre-eminence to freedom of expression when political criticism is being exercised[85]. This requires there to be sufficient and extensive motives to violate freedom of expression, which, in turn, requires criminal courts to consider the author's intent with respect to their messages.

III. EMPIRICAL ANALYSIS OF REGULATORY COMPLIANCE AND SELF-CENSORSHIP IN SOCIAL NETWORKS

III.1. Regulatory compliance, self-censorship and banning expressions on the Internet: measuring the impact of the restrictions

The previously identified increase in judicial intervention and convictions for expressions on social networks restricts freedom of expression[86]. However, this could be justified by the social interests that the criminal justice system aims to protect. The point, as already stated, is that in addition to the questionable extent of criminalization of crimes of expression,[87] it is possible that any criminal law limitation on freedom of expression may actually be extended by the fact that citizens are not able to determine what is permitted or not[88]. This would mean that the normative effect of compliance, either through the threat of punishment or any other mechanism, would significantly damage freedom of expression beyond what is directly restricted.

The empirical study presented below aims to examine the real impact that both private and public regulation of freedom of expression in social networks has on compliance with these rules and on the decision to self-censor and not to publish certain content. From the perspective of regulatory compliance, we believe it is necessary to comprehend the prevalence of offensive and disallowed publications on social networks such as Twitter, and what factors are associated with non-compliance with the rules. To this end, we have analysed variables based on three regulatory compliance perspectives that criminological literature has analysed in detail: deterrence, so-

[84] Cesar Strawberry is a well-known singer of the rock band "Def con Dos". The Twits for which he was sentenced to one year in prison by the Supreme Court were the following: 1. "Esperanza Aguirre's uncomplicated fascism makes me miss even the GRAPOs". Esperanza Aguirre is a Spanish politician from the Popular Party, considered a right-wing party in Spain. Also, GRAPO was a Spanish terrorist group. 2. "Ortega Lara should be kidnapped now". Ortega Lara was a prison officer and Popular Party member who was kidnapped by the terrorist group ETA, and was held hostage for 532 days. 3. "Street Fighter, post-ETA edition: Ortega Lara versus Eduardo Madina". Eduardo Madina is a politician of the Socialist Party, considered a leftist party, who suffered an ETA attack in 2002, which caused him injuries. 4. "Franco, Serrano Suñer, Arias Navarro, Fraga, Blas Piñar If you don't give them what Carrero Blanco got, longevity will always be on their side".Franco, Serrano Suñer, Arias Navarro, Fraga and Blas Piñar were extreme right-wing politicians and supporters of Franco. Carrero Blanco was head of government in the final stage of Franco's dictatorship, and he died as a result of an attack by ETA. 5. "How many should follow Carrero Blanco's flight". 6. "It's almost the King's birthday. (I'm going to give him) a donut-bomb".

[85] This question has already been analysed in MIRÓ LLINARES, F., "Derecho Penal y 140 caracteres...", op. cit.

[86] Perhaps even the mere processing of complaints for certain expressions, even if the cases are subsequently filed, also has an impact on this right.

[87] See MIRÓ LLINARES, F., "La criminalización de conductas offensivas...", ob. cit. Similarly, with different approaches but in a very similar critical sense, MIRÓ LLINARES, F., "*Cometer delitos en 140 caracteres...*", ob. cit., and also Grupo de Estudios de Política Criminal, "*Una propuesta alternativa...*", ob. cit.

[88] This is not difficult to imagine given, for example, the different interpretations by the courts of which expressions may be criminal and which may not. Thus, for example, the well-known case of Councillor Zapata who, until he was freely acquitted,

cial influence, and legitimacy[89]. In short, the deterrence perspective holds that compliance with rules essentially depends on the characteristics of the punishment associated with non-compliance, especially the perceived severity and certainty of the punishment[90]. Moreover, this is the assumption behind the legislative decision to criminalize conducts by threatening punishment, especially when the severity of punishment is increased. This hypothesis has also been questioned by numerous empirical studies in the field of compliance of many types of rule[91]. The social influence perspective, on the other hand, holds that compliance depends on two major social norms[92]: the descriptive norm, which provides the subject with information on acceptable behaviour based on how others behave; and the prescriptive norm, which indicates how the reference group will morally judge the subject's behaviour[93]. Finally, the legitimacy perspective holds that compliance with a rule depends on, among other variables, the agent's moral judgment with respect to the conduct. In the sense that the worse the subject believes a certain behaviour is, the less he will do it and consequently the more he will comply[94]. In addition to moral judgement, in the case of offen-

had to go through various judicial procedures, or the case of Cassandra Vera, who was first convicted and then acquitted. As LASCURAÍN SÁNCHEZ, J. A., states, "Everything at once: limitation of expression and lack of protection of one's honour", in *Revista Jurídica Universidad Autónoma de Madrid*, no. 26, 2017, p.125: "with new criminal laws or new interpretations of the laws or the Constitution we have the feeling that in recent times these limits have been brought closer and that we feel less free, more dissuaded, more discouraged, to express our opinions. We can be accused for making jokes (the Cassandra case), or for making acidic, cruel but political comments (the César Strawberry case); we can be accused for expressing our opposition to the existence or recognition of transsexuality (the Transvestite Bus case); we can be accused for singing an anthem (the whistle case); we can be condemned for burning a flag or the portrait of a Head of State; our punishment for violently preventing a political act can be aggravated for reasons linked to the expression, for shouting "Catalanidad es Hispanidad" (Blanquerna case)".

89 On the study of these approaches and their relationship with compliance with certain rules, see MIRÓ LLINARES, F. & BAUTISTA ORTUÑO, R., "¿Por qué cumplimos las normas penales: Sobre la disuasión en materia de seguridad vial", in *Indret: Revista para el Análisis del Derecho*, No. 4, 2013; and, GÓMEZ BELLVÍS, A. B, "Crónica de una ineficacia anunciada: Un estudio sobre los factores asociados al cumplimiento en el ámbito de la propiedad intelectual", in *Indret: Revista para el Análisis del Derecho*, No. 1, 2019. In both articles, the authors elucidate the three approaches, the available empirical evidence and the current literature in terms of regulatory compliance.

90 PATERNOSTER, R., "How Much do we really know about criminal deterrence?", *The Journal of Criminal Law and Criminology*, vol. 100, 2010.

91 Such as traffic rules, especially those referring to alcohol intake and speeding (MIRÓ LLINARES, F., & BAUTISTA ORTUÑO, R., "¿Por qué cumplimos...", ob. cit.); on speeding (Dusek, L, & Traxler, C., "Learning from Law Enforcement", in *cesifo Working Papers*, 8043, 2020); hate through social networks (BAUTISTA ORTUÑO, R., "¿Es eres un ciberhater? Predictors of Violent Communication and Discourse of Hate on the Internet", in *International e-Journal of Criminal Sciences*, No. 12, 2018); non-payment of taxes (HICHEM, K. & ACHEK, I., "The determinants of tax evasion: a literature review", in *International Journal of Law and Management*, vol. 57, 2015); copyright infringement on the Internet (GÓMEZ BELLVÍS, A. B., "Crónica de...", ob. cit.) All of them agree that deterrence variables explain very little of the decision to comply with the rules. In this regard, and on the role that punishment should play in relation to what empirical evidence indicates, see MIRÓ LLINARES, F., "La función de la pena ante el paso empírico del Derecho penal", in *Revista General de Derecho Penal*, No. 27, 2017.

92 KAHAN, D. M., "Social Influence, Social Meaning, and Deterrence", in *Virginia Law Review*, vol. 83, 1997; ROBINSON, P. H., *Principios distributivos del Derecho Penal. A quién debe sancionarse y en qué medida*, Marcial Pons, Madrid, 2012; GAYMARD, S., "Norms in social representations: two studies with French Young drivers", in *The European Journal of Psychology Applied to Legal Context*, no. 2, 2009; CIALDINI, R. B., & GOLDSTEIN, N. J., "Social influence: Compliance and conformity", in *Annual Review of Psychology*, vol. 55, 2004, among many others.

93 CIALDINI, R. B., KALLGREN, C. A., & RENO, R. R., "A Focus Theory of Normative Conduct: A theoretical refinement and reevaluation of the role of norms in human behaviour", in *Advances in Experimental Social Psychology*, vol. 25, 1991.

94 TYLER, T., *Why people obey the law*, Princeton University Press, Oxford, 2006; TYLER, T., "Compliance with Intellectual Property Laws: A Psychological Perspective", en *New York University Journal of International Law and Policy*, vol. 29, 1997.

sive messages, we understand that the perceived offensiveness of the messages can also affect compliance from the perspective of legitimacy, to the extent that a negative moral evaluation of a conduct may take into account how offensive the message may be in general or to the specific person to whom it is directed[95].

But given that it is not only relevant whether and why people comply with rules prohibiting the expression of certain statements but also whether, as a result, people decide to self-censor, we also believe it is important to take into account the variable of self-censorship or, as it is better known in academia, the *chilling effect*[96]. The term *chilling effect* was coined in the United States at the beginning of the 50's[97] by the judge Felix Frankfurter in the United States Supreme Court sentence *Wieman v. Updegraff* (1952)[98] that annulled a loyalty oath on the part of public employees. It was understood that there was the possibility that a large number of people subject to a vague or ambiguous law may not exercise their constitutionally protected freedom of expression for fear of being prosecuted[99]. This highlighted the need to avoid ambiguity and vagueness in laws, and the need for courts to interpret and apply them in a restrictive manner so that there is no unnecessary or disproportionate sacrifice of the freedom which is being restricted[100]. Although the risk of "self-censorship" has usually been discussed in relation to the criminalization of certain declarations by means of laws and court interpretations, a *chilling effect 2.0* could be derived from the control of contents carried out by social networks. As we have seen, this control is sometimes even more restrictive than the legal control and although the consequences are less in terms of the deprivation of rights, the relevance for public communication is salient with respect to the restriction on expression in cyberspace[101].

Given the importance of this effect in providing support for maximising limitations on crimes of expression, there is a striking lack of empirical analysis, especially considering it is descriptive

[95] BAUTISTA ORTUÑO, R., CASTRO-TOLEDO, F. J., PEREA-GARCÍA, J. O., & RODRÍGUEZ-GÓMEZ, N., "May I offend you? An experimental study on perceived offensiveness in online violent communication and hate speech", en *International E-Journal of Criminal Sciences*, no. 12, 2018.

[96] SCHULTZ, D., & VILE, J. R. (EDS.), *The Encyclopedia of Civil Liberties in America*, Volume One, Sharpe Reference, 2005, p. 161; DE DOMINGO PÉREZ, T., "La argumentación jurídica en el ámbito de los derechos fundamentales: en torno al denominado "chilling effect" o "efecto desaliento", in *Revista de Estudios Políticos*, No. 122, 2003; CUERDA ARNAU, M. L, "Proporcionalidad penal y libertad de expresión: la función dogmática del efecto de desaliento", in *Revista General de Derecho Penal*, no. 8, 2007; BEA, D. C., "La doctrina del efecto desaliento como punto de conexión entre el Derecho penal y los derechos fundamentales", in *Cuadernos Electrónicos de Filosofía del Derecho*, no. 41, 2019; among others.

[97] Although, according to the encyclopaedia, the phenomenon itself is much older.

[98] The sentence is available at: https://caselaw.findlaw.com/us-supreme-court/344/183.html

[99] SCHAUER, F., "Fear, Risk and the First Amendment: Unraveling the "Chilling Effect", in *Boston University Law Review*, vol. 58, 1978. On the other hand, while it is true that when we refer to the chilling effect we do so in the context of freedom of expression, it is also generally used to refer only to the phenomenon in which as a consequence of a rule the citizenry is dissuaded from engaging in some behaviour. Thus, for example, the study by CANES-WRONE, B., & DORF, M. C., "Measuring the Chilling Effect", *New York University Law Review*, vol. 90, 2015 in which they attempt to measure the effects of certain abortion laws on abortion behaviours.

[100] STC No. 88/2003, of 19 May.

[101] Especially with regard to social networks, where many users make use of political expression (BRODE, L., VRAGA, E., K., BORAH, P., & SHAH, D. V., "A New Space for Political Behavior: Political Social Networking and its Democratic Consequences", in *Journal of Computer-Mediated Communication*, vol. 19, 2014).

rather than a normative issue[102]. Some empirical studies have attempted to analyse how people behave in social networks, if they self-censor,[103] to what extent they do so and why, especially through the prism of the *Spiral of Silence Theory* developed by Noelle-Neumann[104]. These studies suggest that citizens self-censor if they believe their opinions are in conflict with the dominant positions, and that indeed, the decision not to publish content on social networks is related to social influence[105].

However, there is a lack of research that analyses the factors that influence the decision to publish offensive content, as well as research that determines the impact of regulations on the decision to express opinions protected by the right to freedom of expression.

III.2. Empirical study
3.2.1. Objectives and hypothesis

The present study has two general objectives: on the one hand, to examine the prevalence of messages published by our sample that are offensive and contrary to some of Twitter's policies, and in this way, to analyse the factors associated with noncompliance of these rules. On the other hand, to descriptively evaluate the prevalence of self-censorship and its characteristics within our sample.

These general objectives are further delimited into the following specific objectives:

1. To examine the sample's perception of the freedom of expression in Spain.

2. To analyse the prevalence of offensive messages made on social networks.

3. To assess factors associated with non-compliance with expression rules.

4. To examine the prevalence of self-censor-

[102] TWONEND, J., "Freedom of expression and the chillin effect", in *The Routledge Companion to Media an Human Rights*, 2017. Penney, J. W., "Internet surveillance, regulation, and chilling effects online: a comparative case study", en *Internet Policy Review. Journal on internet regulation*, vol. 6, 2017.

[103] DAS, S., & KRAMER, A., "Self-Censorship on Facebook", Available at: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/viewFile/6093/6350, used sample of 3.9 million Facebook users to show that 71% self-censored at what they call "the last minute", i.e. they analyse users who decide not to post something at the last minute, and also found that those users who had a specific audience self-censored more.

[104] NOELLE-NEUMANN, E., *The Spiral of Silence. Public opinion—our social skin*, The University of Chicago Press, 1993.

[105] KWON, K. H., MOON, S. I., & STEFANONE, A., "Unspeaking on Facebook? Testing network effects on self-censorship of political experssions in social network sites", ¸en Qual Quant, Springer, 2014; HOFFMANN, C. P., & LUTZ, C., "Spiral of Silence 2.9: Political Self-Censorship among Young Facebook Users", 2017. Available at: https://www.researchgate.net/profile/Christoph_Lutz/publication/318475350_Spiral_of_Silence_20_Political_Self-Censorship_among_Young_Facebook_Users/links/59d77016458515db19cb99e6/Spiral-of-Silence-20-Political-Self-Censorship-among-Young-Facebook-Users.pdf Although also related to other variables, as shown by HAYES, A. F., SCHEUFELE, D. A., & HUGE, M. E., "Nonparticipation as Self-Censorship: Publicly Observable Political Activity in a Polarized Opinion Climate", in *Political Behavior*, vol. 28, núm.3, 2006. Similarly, Hayes explains that "after more than 3 decades of research on the spiral of silence, the evidence supporting it is mixed. Although its basic tenets are logically sound and have a certain "intuitive truth", we know that opinion expression in public contexts is multiply determined. Decisions to express an opinion can be based in part by variables central to the theory, such as perceived congruence between one's opinion and perceived public opinion, or a person's fear of social isolation. But decisions to speak out can also be influence by such factors as a person's knowledge about the topic, interest in political matters or public affairs, importance or salience of the topic to the person, confidence in the 'correctness' of one's opinion, the extremity of one's opinion, communication apprehension and shyness, and the extent to which one's opinion is based on moral principle" (HAYES, A. F., "Exploring the Forms of Self-Censorship: on the Spiral of Silence an the Use of Opinion Expression Avoidance Strategies", in *Journal of Communication*, vol. 57, 2007, p. 786).

ship within the sample and the topics that are self-censored.

5. To analyse the reasons for self-censorship.

To achieve the above objectives, the following hypothesis have been formulated. In relation to compliance with the rules:

Based on the deterrence perspective:

H1(a). Greater perceived punishment for transgression of the rule is associated with increased compliance.

H1(b). Higher perceived probability of being sanctioned is associated with increased compliance.

Based on the social influence perspective:

H2(a). Greater social disapproval of the behaviour is associated with increased compliance.

H2 (b). Increased compliance with the rule in the person's reference group is associated with increased compliance.

Based on the legitimacy perspective:

H3(a). A more positive judgement of the conduct is associated with increased compliance.

H3(b). Greater perceived offensiveness of the message is associated with increased compliance

In contrast, the following hypotheses have been formulated in relation to self-censorship:

H4(a). The prevalence of self-censorship in the sample will be high

H4(b) The topics on which they will exercise self-censorship will be mainly political topics

H4(c) The reason for exercising self-censorship will be related to the possibility of being sanctioned[106]

3.2.2. Method

3.2.2.1. Sample

The sample (*N=443*) is composed of 53% men and 47% women, with a mean age of 34 years (*SD=12.21*). 86.7% of the participants are educated to university level, and with respect to political ideology, the sample was on average on the left of the political spectrum. In this sense, on a scale from 1 to 7, where 1=extreme left and 7=extreme right, the average ideological position is 3.14 (*SD=1.18*).

3.2.2.2. Design, variables, procedure and instrument

A non-experimental design was used to achieve the objectives of this study and to test the above-mentioned hypotheses.

The dependent variable is compliance with four specific usage policies regarding expressions not allowed on Twitter. In this sense, we have taken into account the rules prohibiting threats, glorification of terrorism, discrimination and harassment. We have considered these four insofar as this type of violent communication may coincide with the criminal law prohibitions. In addition, we have added expressions of bad taste, since these can be offensive and offend individual or collective sensibilities, and they are also the most prevalent offensive messages according to some studies[107]. With regard to independent variables, we have taken into account the variables from the deterrence, social influence and legitimacy perspectives. In the latter we have included not only

[106] These three hypotheses have been established on the basis of arguments concerning the possibility that the *chilling effect* may be caused by the rules in general and, in particular, by criminal law. Furthermore, the hypotheses are based on the fact academics believe that the criminalization of eminently political messages may lead to citizens ceasing to express this type of political expression, and that the reason, therefore, for exercising this self-censorship derives from the corresponding rules and sanctions.

[107] Miró Llinares, F., "Taxonomy..." Ob. Cit; Bautista-Ortuño, R., "Are you a cyberhater...", ob. cit.

the moral judgment of the behaviour, but also the perceived offensiveness of the messages. Moreover, we controlled for other variables, such as: a) sociodemographic variables; b) variables related to the perception of the quality of freedom of expression; c) variables related to the perception of the prevalence of censorship; d) variables related to self-censorship, amongst others. Further information on the variables can be found in ANNEX 1.

An *ad hoc* questionnaire was developed to measure all the above variables. Design of the questionnaire was the result of, on the one hand, adapting previous instruments designed to measure regulatory compliance[108], and, on the other hand, several meetings between experts in criminal law and methodology.

Table 4. Descriptive table of the means of the variables of perceived quality of freedom of expression, perceived censorship, and knowledge of the limits of freedom of expression

Variable	Leve	*N*	*%*	*M*	*DT*	*Min*	*Max*
Perceived quality of freedom of expression	Spain	443	100	2.97	0.96	2	5
	Social networking	443	100	2.93	1.144	1	5
	Twitter	443	100	3.01	1.234	1	5
Perceived Censorship on Twitter	From social networks to users	443	100	3	1.048	1	5
	Perception of objectionable content on Twitter	443	100	3.11	1.285	1	5
Knowledge Law	No	78	17.6	-	-	-	-
	Yes	342	77.2	-	-	-	-
	Don't know/no answer	23	5.2	-	-	-	-
Political Knowledge Twitter	No	212	47.9	-	-	-	-
	Yes	191	43.1	-	-	-	-
	Don't know/no answer	40	9	-	-	-	-

[108] Miró Llinares, F., and Bautista Ortuño, R., ¿Por qué cumplimos...", ob. cit.; Gómez Bellvís, A. B., "Crónica...", ob. cit. In order to operationalize the perceived offensiveness of the messages, the scale used was from Bautista Ortuño, R., Castro-Toledo, F. J., Perea-García, J. O., & Rodríguez-Gómez, N., "'May I offend you...". Ob. cit.

The questionnaire was administered through the social network Twitter. Google's free survey system was used to develop the survey. The criteria for inclusion in the sample were: 1) be a Twitter user and reside in Spain; 2) be at least 13 years old and, 3) speak Spanish. Research was conducted from 17/02/2020 to 28/02/2020, inclusively. Finally, randomization of the sample was carried out through the program *Sublime Text*, which unified the three links corresponding to the questionnaires.

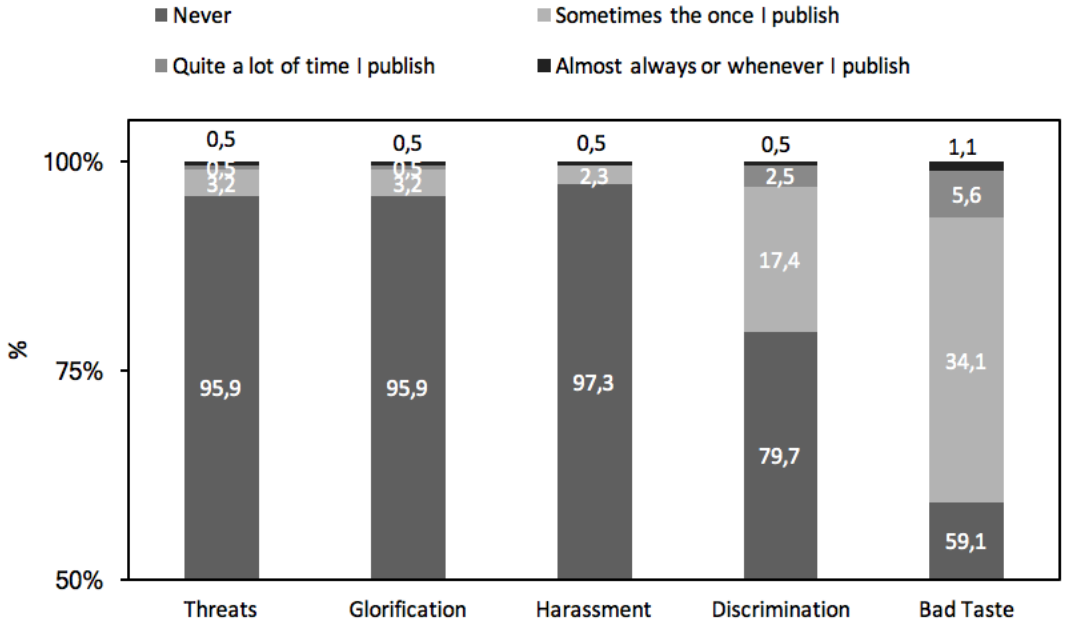

Figure 5. Prevalence of publication of offensive content (%)

IV. RESULTS

a) Perception of the quality of freedom of expression, perceived censorship in social networks and knowledge of the limits of freedom of expression

In relation to the variable "perception of the quality of freedom of expression", three different elements were surveyed. With reference to the quality of this right in general in our country, the average in the sample is almost 3 (M=2.97; SD=0.96), which means that for our participants it is neither good nor bad. This result is striking since, as we have analysed above, our country is among the freest in terms of freedom of expression according to international indicators. The same applies to the perception of freedom of expression on social networks in general, and on Twitter in particular.

With regard to the perception of censorship on social networks such as Twitter, the sample average indicates that Twitter does not censors users a lot or a little and respondents also consider there to be neither much nor little censurable content on social networks such as Twitter (*M=3; SD=1.04*; and, *M=3.11; SD=1.2*, respectively).

Finally, regarding the sample's knowledge about the limits of freedom of expression according to the law, 77.2% admit that they do know what the legal limits are. However, regarding the knowledge of Twitter policies in relation to what can and cannot be expressed, about 50% of the

sample says they do not know, as shown in Table 4.

b) Informed compliance

As regards the variable "informed compliance", that is, the variable through which we try to obtain information about the type of messages that the sample publishes and that go against Twitter policies, Figure 5 shows the prevalence is very low except in the category of messages of bad taste where 34.1% of the sample say they do it some of the times they publish, 5.6% say they do it quite often and 1.1% say they do it almost always or whenever they publish. Likewise, with regard to messages that can be understood as discriminatory, although nearly 80% of the sample say they never do it, 17.4% report doing it some of the times they publish, 2.5% quite a few times and 0.5% almost always or always.

c) Variables associated with compliance

In this study, we analysed variables that literature on three compliance theories has shown can be related to compliance: deterrence, social influence, and legitimacy. From deterrence theory, we have taken into account perceived severity and certainty. In terms of severity, as depicted in Figure 6, responses vary considerably.

If we take into account those messages that harm some type of interest and whose prohibition also coincides, albeit abstractly, with the criminal law, such as threats, glorification, harassment or discriminatory messages, very few people believe that no formal sanction would be applied to this type of message. However, about 50% in each category consider that the maximum sanction they could receive for publishing this type of message would be a sanction handed down by the social network, that is, either withdrawal of the message or removal of the user's account. The rest, however, consider that the formal sanction they would receive for making this type of expression would be a fine, except in the category of glorification of terrorism, where 28.9% of the sample believe that they would receive a prison sanction. A different scenario can be found with

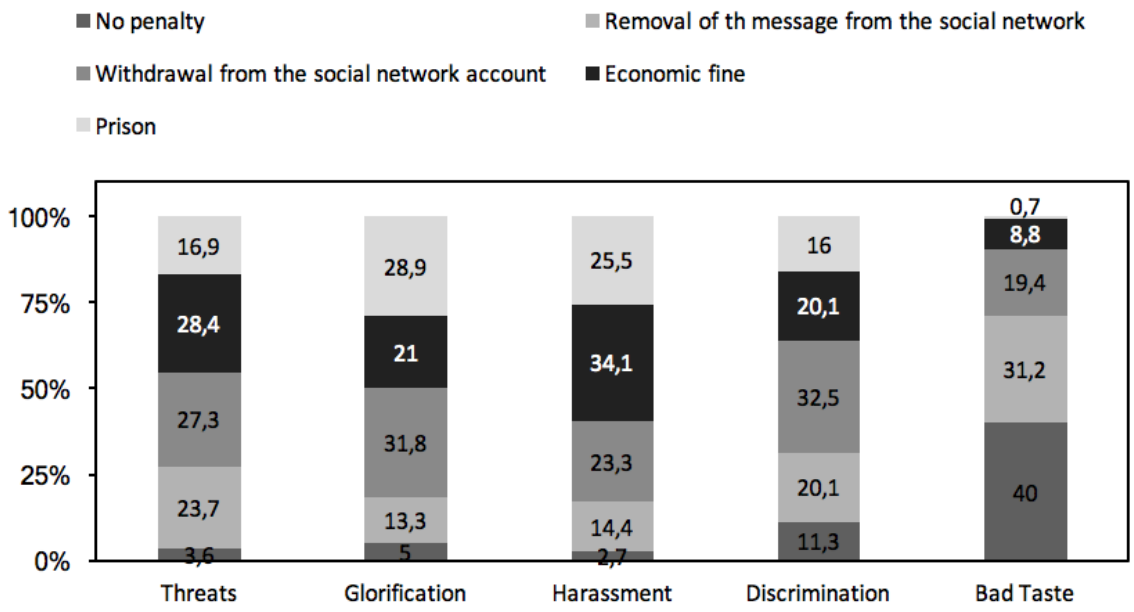

Figure 6. Description of the perceived severity of sanctions for publication of messages on social networks

respect to messages considered bad taste, where 40% of the sample believe that they would not receive any type of sanction; 31.2% considered that the message could be deleted from the social network; 19.4% that their account could be removed; 8.8% that they could receive a fine; and, only 0.7% thinks that they could be punished with imprisonment. Perhaps the likely punishments are believed to be not very severe because these are the most common type of messages that the sample reports publishing on the social network.

With regard to the variable perceived certainty, as detailed in Table 5, on average the sample does not know if they could be sanctioned in all categories. That is, they report a considerable degree of uncertainty regarding the probability of being pu-

Table 5. Descriptive summary of compliance variables and perceived offensiveness

Variable	Conduct	*M*	*SD*	*Min*	*Max*
Certainty	Threats	3,11	1,168	1	5
	Glorification	3,5	1,315	1	5
	Harassment	3,19	1,219	1	5
	Discrimination	3,03	1,263	1	5
	Bad taste	2,43	1,172	1	5
Descriptive standard	Threats	1,6	0,683	1	3
	Glorification	1,57	0,697	1	4
	Harassment	1,64	0,762	1	4
	Discrimination	2,12	0,898	1	4
	Bad taste	2,51	0,891	1	4
Prescriptive standard	Threats	1,41	0,589	1	4
	Glorification	1,28	0,557	1	4
	Harassment	1,33	0,597	1	5
	Discrimination	1,56	0,789	1	5
	Bad taste	2,17	0,948	1	5
Moral judgement	Threats	4,58	0,807	1	5
	Glorification	4,58	0,889	1	5
	Harassment	4,65	0,823	1	5
	Discrimination	4,42	0,977	1	5
	Bad taste	3,44	1,239	1	5
Perceived offensiveness	Threats	4,4	0,88	1	5
	Glorification	4,4	0,94	1	5
	Harassment	4,55	0,78	1	5
	Discrimination	4,35	0,97	1	5
	Bad taste	3,37	1,28	1	5

nished for publishing these messages. Certainty is only a slightly lower in the category of messages that are of bad taste.

With reference to social influence theory, we have taken into account both the descriptive norm, i.e., the extent to which participants believe people publish these expressions; and the prescriptive norm, i.e., the extent to which they believe that their reference group would pass a positive or negative moral judgment if they were aware that they had published these expressions.

In the case of the descriptive norm, on average the sample reports that very few people close to them publish this type of expressions. The average is slightly higher in the category of expressions that are of bad taste. As for the prescriptive norm, that is, moral judgment or disapproval from their reference group, on average the sample reports that people around them would consider it very bad if they published this type of expression.

Finally, with respect to the variable "moral judgment" that forms part of the legitimacy perspective, as can be seen in Table 5, all categories of messages are, in moral terms, close to the statement "totally wrong", except for the messages in bad taste that on average score morally better than the rest but are still closer to a negative judgment than a positive judgment.

The perception of offensiveness or "harm" that this type of message could cause to those targeted was also taken into account. Thus, regarding the offensiveness of the message we can see that on average all categories are quite close to 5, except, again, messages that are of bad taste.

All this information has been included in Table 5.

Having carried out the descriptive analyses of the previous variables, a bivariate analysis was conducted between the different behaviours evaluated by the variable "informed compliance" with its respective variables from deterrence perspective (perceived severity and certainty), from social influence (descriptive and prescriptive norm), and from legitimacy (moral judgement and the variable operationalized as "perceived offensiveness"). Due to the ordinal nature of the different variables, Spearman correlations were conducted.

As shown in Table 6, for all expressions analyzed, significant relationships were found between the different types of behavior and the descriptive norm, prescriptive norm, moral judgment and perceived offensiveness. In this sense, severity only correlates significantly and negatively with messages that are of bad taste, i.e. the lower the perceived severity of punishment, the greater the publication of messages of this type. Certainty, on the other hand, was not significantly correlated with any of the behaviors analyzed. These data imply that we should reject the hypotheses formulated based on the deterrence perspective (*H1(a). Greater perceived punishment for transgression of the rule is associated with increased compliance. H1(b). Higher perceived probability of being sanctioned is associated with increased compliance*). In other words, neither severity nor certainty are related to greater compliance with the rules assessed in this paper, at least according to the traditional hypotheses from this perspective.

The descriptive norm is significantly related to all the behaviours, although the correlation is stronger in the case of the messages considered discrimination or bad taste. In this sense, when greater participants' belief that others publish this type of expressions on social networks, the more likely participants of the study are to also publish similar messages. On the other hand, the prescriptive norm, that is, the reference group's moral judgment also correlates significantly with all behaviors. The strongest correlation is found

Table 6. Correlations between the conducts and the variables for normative compliance/perceived offensiveness

	Severity	Certainty	N. Descriptive	N. Prescriptive	J. Moral	Offensiveness
Threats	-0,022	0,019	,147**	,184**	-,154**	-,114*
Glorification	-0,063	0,018	,225**	,251**	-,272**	-,231**
Harassment	-0,021	-0,022	,151**	,107*	-,153**	-,117*
Discrimination	-0,009	-0,016	,287**	,346**	-,287**	-,176**
Bad taste	-,188**	-0,021	,271**	,392**	-,424**	-,339**

** Correlation is significant at 0.01 (bilateral) * Correlation is significant at 0.05 (bilateral)

with respect to expressions considered discriminatory and of bad taste. In other words, the more positive the reference group's moral judgement is about such expressions, the more likely it is that the participants publish these types of messages. In this sense, we must accept the hypotheses put forward regarding the social influence perspective (*H2(a). Greater social disapproval of the behaviour is associated with increased compliance. H2(b). Increased compliance with the rule in the person's reference group is associated with increased compliance*).

Participants' moral judgment about the different types of messages is significantly related to their dissemination. The strongest correlation was found with respect to expressions of bad taste, discriminatory messages and those that glorify terrorism. In this way, the more negative the particpants' moral judgment is regarding these types of expression, the less likely it is that they publish these messages. Finally, the perceived offensiveness was also correlated with these behaviors, especially with respect to expressions that are of bad taste, that discriminate or that glorify terrorism. That is, the more offensive the participants believe the message is, the less this type of declaration is made by the sample. This leads us to accept the hypotheses formulated regarding the legitimacy perspective (*H3(a). A more positive judgement of the conduct is associated with increased compliance. H3(b). Greater perceived offensiveness of the message is associated with increased compliance*).

d) Self-censorship: scope and reasons

One of the variables of special interest for this study was informed self-censorship. In this regard, Table 7 shows the results for whether participants had decided not to express their opinion on any type of topic on social networks such as Twitter in the last 12 months which they would have liked to express. Only 19.6% said they had never self-censored; however, 47% of the sample said they had sometimes censored themselves, 26.2% quite a few times and 7.2% said they had done so almost always or always. These findings are particularly notable because they imply that more than 80% of the participants in this study report having stopped themselves from expressing their opinions on social networks sometime in the last year, even though they would have liked to do so. In this sense, we accept hypothesis *H4(a). The prevalence of self-censorship in the sample will be high.*

The topics on which the participants say they have avoided giving an opinion, i.e., they have self-censored, can be seen in Table 8. The majority of the sample stopped giving their opinion on political issues. Perhaps this is the most salient

Table 7. Informed self-censorship

Self-censorship (last 12 months)	N	%
Never	87	19,6
Some of the times I publish	208	47
Quite a lot of the times I publish	116	26,2
Almost always or whenever I publish	32	7,2

finding, since freedom of expression is particularly relevant when the content of what is expressed is political. Therefore, we accept hypothesis *H4(b) The topics on which they will exercise self-censorship will be mainly political topics.*

As for the reasons why the sample self-censors, as detailed in Table 9, from the various options available to the participants the reason "I have stopped expressing my opinions because I believe that what I wanted to express could offend other people" stands out as 80% of the sample reports that this is one of the reasons why they

Table 8. Topics on which opinions have been avoided

Issues avoided on social networks		N	%
I avoid giving opinions on political issues	Yes	284	64,1
	No	158	35,7
I avoid giving opinions on gender issues (gender violence, feminism, machismo)	Yes	30	6,8
	No	413	93,2
I avoid giving opinions on religious matters	Yes	43	9,7
	No	400	90,3
I avoid commenting on economic issues	Yes	11	2,5
	No	432	90,3
I avoid giving opinions on issues of nationalism in Catalonia	Yes	6	1,4
	No	437	98,6
I avoid commenting on criminal justice issues	Yes	6	1,4
	No	437	1,4
I avoid talking about other subjects (e.g. love, science, corruption, pollution, education, etc.)	Yes	48	10,8
	No	395	89,2

Table 9. Reasons for self-censorship

Reasons not to give your opinion on social networks.		*N*	%
I have stopped expressing my opinions because I thought they might delete or block the message.	Yes	41	9,3
	No	402	95
I stopped expressing my opinions because I thought my social network account might be deleted.	Yes	22	5
	No	421	92,6
I stopped expressing my opinions because I thought I might be punished for a crime.	Yes	33	7,4
	No	410	92,6
I have stopped expressing my opinions because I thought my friends, family or close people would not like my opinion.	Yes	77	17,4
	No	366	82,6
I stopped expressing my opinions because I thought others (third parties) would not like my opinion.	Yes	111	25,1
	No	332	74,9
I have stopped expressing my opinions because I believe that what I wanted to express could offend other people.	Yes	356	80,4
	No	87	19,6
I stopped expressing my opinions because I believed that what I was about to express should not be made public.	Yes	111	25,1
	No	332	74,9
Other reasons (e.g. avoiding a confrontation, not worth getting into conflicts, fear of being attacked, fear of future repercussions, etc.)	Yes	103	23,3
	No	340	76,6

have stopped expressing their opinion. Likewise, 25.1% revealed that they stopped expressing their opinions because they thought their opinion would not please others or because they believed their opinion should not be made public, and 17.4% because they believed that it would not please the people around them. In other words, 42.5% of the sample stopped expressing their opinions because they believed that others would not like what they say. But, perhaps the most relevant data for the purpose of verifying the possible effect the rules have on self-censorship, which we have called the *chilling effect 2.0*, is that according to the data in this study the possibility of deleting or blocking the message or the social network account, or even the corresponding punishment if their expression consitutes a criminal offence is not a relevant reason for self-censorship. Consequently, we must reject hypothesis *H4(c) The reason for exercising self-censorship will be related to the possibility of being sanctioned.*

V. Discussion and limitations

In the present research we aimed to analyse the state of freedom of expression in Spain in light of the existing concern from international organizations and experts who report that state control has increasingly limited what can be expressed freely, and considering that social networks seem to have had some influence on the restrictions. Analysis of the legislative evolution from 1995 to the present day regarding the most problematic crimes of expression served to verify that the successive Criminal Code reforms have increased the scope of what is punishable and therefore the catalogue of declarations that cannot be expressed. This is especially lucid with regard to the crimes of hate speech and glorification of terrorism and humiliation of its victims. With the aim of analysing whether this criminal law reduction in free expression had translated into a reduction in reality, we carried out an exploratory analysis of sentences to analyse the judicial application of these types of criminal offences. It was noted that the increase in sentences in recent years is particularly striking for the crimes of glorification of terrorism and hate speech, and we have also seen that social networks and the Internet have had an impact on these sentences, insofar as the majority of expressions that have been prosecuted as of 2014 occur in cyberspace. This could give rise to the idea that social networks have increased or precipitated the commission of this type of crime, but another possible hypothesis that could explain this trend is that there has been an increase in the control of what is expressed. It is important to point out at this point that the data on sentences analysed in this paper should be taken and interpreted with caution. Due to the inaccessibility of the data, we have not analysed all indicators of judicial application, such as, for example, judicial investigations that have been carried out for the possible commission of these crimes but that have not ended in a trial. Without doubt, these data would help us understand the real dimension of the issue and which, we intuit, is much more extensive than what we have been able to evaluate here[109]. On the other hand, the sample of sentences analysed here does not represent all sentences. Rather, they are a sample obtained based on inclusion criteria from the Aranzadi database that does not include all the sentences passed by the Courts. Likewise, in the present paper we have focused only on crimes of expression in the Criminal Code and have not analysed the effects that the Citizen Security Law, popularly known as the "Gag Law", may have had on freedom of expression. Despite these limitations, what we can say from the analysis of sentences carried out here is that in Spain the popularization of social networks turned two of the crimes that applied very infrequently into a significant source of criminal proce-

[109] Especially if we take into account the data provided by the latest reports by the State Public Prosecutor's Office. Thus, for example, according to the 2019 Report by the State Prosecutor's Office, which analyses the data for the year 2018, in relation to article 510 on hate speech, there were 117 judicial proceedings for incitement (article 510.1) that the State Public Prosecutor's Office is following up, 56 investigative proceedings opened in the State Prosecutor's Office; 15 indictments and 9 sentences. For humiliation or justification of this type of crime (article 510.2) there were 316 judicial proceedings monitored by the State Public Prosecutor's Office, 54 investigative proceedings opened by the State Public Prosecutor's Office, 57 indictments filed, and 23 sentences handed down. The 2018 report of the Attorney General's Office, which evaluates 2017, indicates that for hate speech (article 510.1) there were 89 judicial proceedings monitored by the State Public Prosecutor's Office; 101 investigative proceedings initiated by the State Public Prosecutor's Office, 14 indictments and 19 sentences. For humiliation or justification (article 510.2) there were 220 judicial proceedings monitored by the State Public Prosecutor's Office, 75 investigative proceedings, 52 indictments and 19 sentences.

edings and charges. Furthermore, the judicial interpretation of the first offenses of this type was more expansive and wide-ranging, in our opinion too much so, than the interpretations that came later. However, as we have pointed out, it is possible that the trend in criminalization of this type of expression will change as a result of recent interpretations by our Constitutional Court, especially with regard to the César Strawberry case. In this ruling, the Court makes clear it is necessary to provide due arguments for limiting freedom of expression, particularly with regard to those expressions that, although unpleasant, constitute political criticism. Perhaps what this will produce is a Strawberry effect[110], an effect quite contrary to the *chilling effect*, a belief that now it is possible to affirm everything in social networks as long as it constitutes political or ideological criticism. Or, simply, the effect might be increased belief that in a democratic country like ours no one is going to be convicted for defending even the most abject ideas. In any case, we will have to wait to be able to observe both this new jurisprudential tendency that we intuit will take place, as well as the effect that it may have on what citizens can or cannot express on social networks.

The other major objective of the present research was to use empirical methodologies to move towards analysis of the effect that all this regulation might have had on freedom of expression, both in terms of compliance with private and public rules regulating the publication of content and in terms of self-censorship. The results show that the citizens surveyed have neither a particularly positive nor a particularly negative perception of this issue, which, in a state that is supposed to guarantee freedom of expression, cannot be assessed positively. As for the prevalence of compliance, as expected, respondents rarely acknowledge publishing messages that might be against Twitter policies or the law, although they do admit to posting expressions that might be considered bad taste[111]. The results of the analysis of the factors associated with compliance factor are more relevant. These show how, in line with what we have analysed for other criminal laws[112], variables derived from deterrence do not explain compliance[113]. On the other hand, non-compliance with the rules does seem to be related to the perception of what others do and to a moral judgement of the legitimacy of the conduct[114]. Similar results can be found in the study carried out by Bautista Ortuño[115].

These findings lead us to two important conclusions that we will first express separately and then analyse together in order to fully understand

[110] In fact, there are already headlines such as "TC raises ceiling on freedom of expression by overturning Cesar Strawberry's sentence" https://elpais.com/politica/2020/02/25/actualidad/1582639252_567110.html

[111] This would confirm the data in the study carried out by MIRÓ LLINARES, F., "Taxonomía...", ob. cit., on a sample of more than 250,000 tweets and in which the prevalence of violent messages was particularly low. It would also confirm the results of BAUTISTA ORTUÑO, R., "¿Es eres un ciberhater?...", ob. cit., that in a sample of 1502 Internet users, the prevalence of violent communication showed that messages related to incitement to violence or threats were the least published by users, while messages offending collective sensibilities were published more frequently than any other type of message evaluated in the study.

[112] MIRÓ LLINARES, F., & BAUTISTA ORTUÑO, R., "¿Por qué cumplimos...", ob. cit.; GÓMEZ BELLVÍS, A. B., "Crónica...", ob. cit.

[113] Findings that are consistent with the available literature on the ineffectiveness of using formal sanctions for the prevention of certain behaviors (See MIRÓ LLINARES, F., & BAUTISTA ORTUÑO, R., ¿Por qué cumplimos...", ob. cit.; GÓMEZ BELLVÍS, A. B., "Crónica...", ob. cit.; TYLER, T., "Legitimacy and criminal justice: The benefits of self-regulation", in *Ohio State Journal of Criminal Law*, vol. 7, 2009, among many others).

[114] MIRÓ LLINARES, F., & BAUTISTA ORTUÑO, R., ¿Por qué cumplimos...", ob. cit.; GÓMEZ BELLVÍS, A. B., "Crónica...", ob. cit.

[115] BAUTISTA ORTUÑO, R., "Eres un ciberhater...", Ob. cit.

how, in our opinion, the findings relate to the main focus of this research: the impact that criminal law has on citizens' free expression. The first is that, given the variables related to deterrence do not explain compliance with the rules, the increased criminal repression observed in courts' practices will not necessarily lead to a reduction in what citizens decide to publicly express or not. Increasing punishment for a type of offensive declaration or increasing the certainty that a particular expression might be sanctioned does not seem to determine whether or not it is expressed by citizens. The second conclusion is that serious punishment of acts such as those analysed in this study and punished via crimes of expression seems to be counterproductive to the effects of avoiding these declarations given that in our sample, the lower the perceived severity, the greater the emission of messages of bad taste. This is coherent with the explanatory power of the participants' moral judgment for normative compliance[116]. It is also important in order to evaluate how the law and the regulation of contents on the Internet affects the decision to publish contents: those people who believe they have the right to freely express what the law considers should not be expressed, will continue to do so, while those who consider that publishing these expressions is morally unacceptable or that it can offend others will decide not to do so.

Does this imply that criminal repression of certain types of expression does not affect freedom of expression or, rather, the decision of citizens to express themselves freely? Not really. What it does indicate is that criminal law and the rules that regulate content on the Internet do not seem to be able to change the conduct of those who decide to make expressions contrary to those rules, probably because of the low percevied certainty of being caught. Yet, it is possible that they have affected those who have decided not to express themselves in that sense, perhaps because they consider the rule to be morally adequate and thus, if it were changed and expanded, they may then consider that what is now permitted to be expressed is also "morally adequate". However, these results do show that when the rules go beyond what the citizens themselves consider morally adequate, these citizens decide not to submit to the restriction of freedom of expression. According to the results of the study, citizens who believe they have the right to express something in bad taste decide to do so even if they perceive it to be illegal or that it could lead to a restriction of content by a social network. Does this imply that the law does not affect freedom of expression? We believe that it should be expressed in the opposite sense: it implies that the law is not able to prevent the expression it wants to avoid, but it does end up sanctioning these acts and, therefore, we could say that it makes the expression of these ideas less free, even though it does not manage to stop them from being made public. In fact, there is no doubt that judicial intervention makes these ideas more notorious. But that does not mean that those who express them and are prosecuted or punished for them are, and feel, free. Criminal law, by criminalizing certain expressions that many citizens do not believe should be criminalized, does not achieve its objective of preventing them but erodes the perceived right of citizens to express these ideas freely.

This is totally consistent with the findings on the issue of self-censorship in social networks. Our aim was to examine whether this could materialize with the introduction of rules, insofar as rules can lead to self-censorship of free expression according to academic literature and even the courts. Both these sources warn there is a risk of discouraging people from expressing their ideas or opinions by

[116] MIRÓ LLINARES, F., "La función de la pena...", ob. cit.

introducing rules or laws that restrict freedom of expression and judicial application of the laws. Our data do not indicate that this is the case, though they do show a concerning level of self-censorship. A large proportion of our sample reports having engaged in self-censorship at some time, and 64 per cent acknowledge that politics is the issue which they self-censor. This should be of concern, since freedom of expression makes particular sense for political issues and political criticism, and it helps shape public and therefore democratic debate. With regard to the reasons why the sample decided to self-censor, it is particularly noteworthy that 80.4% reported having refrained from expressing opinions because they believed what they wanted to express could offend other people. It is also noteworthy that nearly 20% of the sample refrained from expressing their opinions because they believed that people around them would not like it. Similarly, 25% say they self-censor because they believe that what they want to express would not be to the liking of others. In other words, more than 40% of the sample decides to self-censor as a consequence of social influence, based on what others may think of the person exercising freedom of expression. This is consistent with the studies that test the *Spiral of Silence Theory*, which show that self-censorship (especially on political issues) is exercised when one's opinion is aligned with the dominant positions[117]. Finally, more than 20% of the sample says that they self-censor for other reasons, such as to avoid confrontation or because they think it is not worth entering into conflict, etc. Similar reasons for exercising self-censorship on Facebook were found by Sleeper, Balebako, Das, McConahy, Wiese & Cranor[118].

Despite the relevance of the results of the present study, we recognize that they should be interpreted with caution due to their limitations. We believe that some items should be that some items should be rethought[119]. It is necessary to carry out more studies on this issue in order to be able to compare results and to advance knowledge on the topic. Similarly, this is not a representative sample of the population, and we must highlight a high level of bias, especially with regard to the education level of the participants, as the majority report having higher education. We consider, in any case, that research on freedom of expression should advance and, especially, the empirical kind, which should provide us with information on its state of health. This information should not only come from analysis of state control conducted through the enactment of laws and their application, but also of that part of freedom that is no longer exercised because citizens prefer not to give their opinion on important issues such as politics. Furthermore, information and transpa-

[117] GEARHART, S., & ZHANG, W., "Was It Something I said?" "No, It Was Something You Posted! A Study of the Spiral of Silence Theory in Social Media Contexts", in *Cyberpsychology, Behavior, and Social Networking*, vol. 18, núm. 4, 2015; CHEN, H. T., "Spiral of silence on social media and the moderating role of disagreement and publicness in the network: Analyzing expressive and withdrawal behaviors", in New Media & Society, vol. 20, 2018; STOYCHEFF, E., "Under Surveillance: Examining Facebook's Spiral of Silence Effects in the Wake of NSA Internet Monitoring", in *Journalism & Mass Communication Quarterly*, vol. 93, 2016, among others.

[118] Although the reasons for this could vary depending on the type of content which the sample in this study decided to self-censor; for example, how the participants want others to perceive them, which is a reason that we have not considered in this study (SLEEPER, M., BALEBAKO, R., DAS, S., MCCONAHY, A. L., WIESE, J., & CRANOR, L. F., "The Post that Wasn't: Exploring Self-Censorship on Facebook", in Proceedings *of the 2013 conference on Computer supported cooperative work*, 2013.

[119] Thus, for example, we do not know for sure if the perception of the quality of freedom of expression shows these values on average because this is the user's opinion or as a consequence of the operationalisation of these variables. Similarly, given the prevalence of expressions that are of bad taste in the sample, we believe that it would be convenient to operationalize in another way what can be considered as "bad taste".

rency about the private control of freedom of expression or censorship 2.0 is also needed.

But while we wait for research to move forward, perhaps we should take advantage of the notoriety that the issue has gained from the Constitutional Court's ruling on the César Strawberry case. In this sense, we could proceed with the very necessary criminal reform of crimes of expression in terms and return to a situation in which only those expressions that are truly harmful (or offensive in a restricted sense that omits any form of political criticism, as has been defended in other studies[120]) are criminalized. Thus, requesting the repeal of most of the conducts included in the previous Criminal Code that do not adhere to the principles of criminal law in a democratic society[121]. We should also reflect on how to address the relationship between state control and private control of freedom of expression, especially the inconvenience of requiring large companies to be responsible for the content that flows on their platforms and which, in some way, obliges them to adopt restrictive measures that are much broader than the law itself[122]. Perhaps social media should take advantage of the results of empirical research such as those found here, as well as seeking more and better empirical findings. This will help them understand that when content policies, including those that restrict the freedom to express certain ideas, are not aligned with citizens' perceptions of what should and can be expressed, it is less likely that the rules will be followed and there may be less satisfaction with the social network itself[123]. Transparency and community participation are not only an abstract duty derived from a duty to be democratic, but rather they seem to be a real foundation for the legitimacy needed to achieve a freer and healthier society[124].

[120] MIRÓ LLINARES, F., *Cometer delitos...*, ob. cit.; CORRECHER MIRA, J., "¿Fin de la broma? El caso Strawberry y el canon constitucional sobre libertad de expresión aplicado al enaltecimiento del terrorismo", en *Diario La Ley*, núm. 9600, 2020.

[121] GRUPO DE ESTUDIOS DE POLÍTICA CRIMINAL, *Una propuesta alternativa de regulación de los delitos de expresión*, Tirant lo Blanch, Valencia, 2019. Along the same lines, the report prepared by the Platform in defence of the freedom of information together with Access Info Europe, the Federation of Journalists' Unions and the Study Group on Criminal Policy and Research Group "Legal regulation and participation of the digital citizen" of the Complutense University of Madrid for its consideration in the 35th session of the Working Group of the United Nations Human Rights Council, where they essentially advocate for the decriminalization of the majority of expression crimes. Available at: https://www.access-info.org/wp-content/uploads/EPU-Espa%C3%B1a-2019-informe-Final.pdf

[122] JACKSON, B. F., "Censorship and Freedom of Expression in the Age of Facebook", en *New Mexico Law Review*, vol. 44, 2014.

[123] Especially if we consider the fact that each user of these social networks can actually isolate themselves from any offense. They have tools at their disposal such as choosing who to follow or who to keep in their contact network, selecting the type of content they want to appear in their respective accounts, blocking even messages or users to prevent them from communicating with them, etc. As discussed in depth elsewhere (MIRÓ LLINARES, F., " La criminalización...."), one of the requirements that an offense had to meet in order to be criminalized according to legal philosopher JOEL FEINBERG, was "the reasonable avoidability standard", understood as the difficulty that potential unwitting witnesses of the offensive words or declarations may have to avoid being witness. According to this standard, the easier it is for the public to avoid the environment in which the conduct has occurred, the less serious the offense will be, and conversely, the more difficult it is to avoid encountering it, the more serious it will be. Feinberg uses the example of "dirty books", asking who is offended by the content of an obscene book that is on the shelf of a bookstore waiting to be read? In this regard, the author states that "no one has the right to be protected by the State against offensive experiences if he could easily and effectively avoid them without any inconvenient or unreasonable effort" (FEINBERG, J., *Offence to Others: The Moral Limits of the Criminal Law*, vol. 2, Oxford University Press, Oxford, 1986.

[124] As LASCURAÍN expresses, "despite the Internet, we must vindicate political expression, because democracy is included in it" (LASCURAÍN SÁNCHEZ, J. A., "Todo a la vez...", ob. cit.

ANNEX 1

Table 7. Study Variables

Socio-demographic	Sex	What's your gender?	Male; Female
	Age	What's your age?	
	Level of studies	What is your level of education?	Primary school; Secondary school; Vocational training; High school; University studies
	Political spectrum	Where would you place yourself on the political spectrum	If 1=left end and 7=right end
Perceived quality of freedom of expression	In Spain	In general, what do you think is the "state of health" of freedom of expression in Spain?	1=Very good; 2=Good; 3=Neither good nor bad; 4=Bad; 5=Very bad
	On social networks in general.	In general, what do you think is the "state of health" of freedom of expression on social networks?	1=Very good; 2=Good; 3= Neither good nor bad; 4=Bad; 5=Very bad
	On Twitter in particular	In particular, what do you think is the "state of health" of freedom of expression on Twitter?	1=Very good; 2=Good; 3= Neither good nor bad; 4=Bad; 5=Very bad
Perception of censorship in social networks	To what extent do you think social networks such as Twitter censors users' expressions (whatever their opinions may be)?	1=No censorship at all and 5=Lots of censorship	
Perception of objectionable content on social networks	To what extent do you think that social networks like Twitter produce a lot of content and messages that could be objectionable from a legal point of view?	1=Not much objectionable content and 5=A lot of objectionable content	

Awareness of the limits of freedom of expression	In the law	Do you know the limits of freedom of expression from a legal point of view? I mean, do you know what can be expressed or not expressed according to the law?	Yes; No; NS/NC
	From social networks	Do you know the limits to freedom of expression according to the policies of social networks such as Twitter, that is, do you know what can be expressed or not according to the policies of social networks such as Twitter?	Yes; No; NS/NC

Table 8. Continuation of the description of the study variables

Informed compliance with social media standards or policies such as Twitter	Threats	In the last 12 months, have you posted any messages (text, images, videos, etc.) on social networks that could be understood as threatening a person or group of people?	1=Never; 2=Sometimes; 3=Often; 4=Almost always or whenever I publish
	Glorification of terrorism or violence	In the last 12 months, have you posted any messages (text, images, videos, etc.) on social networks that justify or glorify any type of violence, terrorist act or violent extremism?	1=Never; 2=Sometimes; 3=Often; 4=Almost always or whenever I publish
	Harassment	In the last 12 months, have you posted any messages (text, images, videos, etc.) on social networks that could be understood to mean that you were harassing someone, or inciting someone else to do so?	1=Never; 2=Sometimes; 3=Often; 4=Almost always or whenever I publish
	Hate speech	In the last 12 months, have you posted any messages (text, images, videos, etc.) on social networks that were against other people based on their political ideology, race, ethnicity, nationality, sexuality or gender?	1=Never; 2=Sometimes; 3=Often; 4=Almost always or whenever I publish

	Tasteless or offensive	In the last 12 months, have you posted any messages (text, images, videos, etc.) on social networks that, while not entering any of the above, could be considered in bad taste or interpreted as offensive, unpleasant or politically incorrect?	1=Never; 2=Sometimes; 3=Often; 4=Almost always or whenever I publish
Deterrence variables	Perceived severity	In your opinion, what maximum sanction do you think a person could receive for publishing the following content on Twitter? (see informed compliance behaviours)	No penalty; withdrawal of the message from the social network; withdrawal from the account at the social network; financial fine; imprisonment
	Perceived certainty	How likely do you think you would be sanctioned for publishing any of the following expressions (sanction includes removal of content, removal from the social network account, or any other)? (see informed compliance behaviours)	1=I'm pretty sure I wouldn't get sanctioned; 2=I'd be unlikely to get sanctioned; 3=I don't know if I'd get sanctioned; 4=I'd be pretty likely to get sanctioned; 5=I'm pretty sure I'd get sanctioned

Table 9. Continuation of the description of the study variables

Social influence variables	Descriptive standard	In your opinion, how many people around you think publish the following expressions in social networks such as Twitter? (see informed compliance behaviours)	1=No one does it; 2=Few people do it; 3=Many people do it; 4=Everyone does it
	Prescriptive standard	And thinking about the people around you, to what extent do you think they would disapprove of your behaviour or think you did something wrong if they knew you did any of the following on social networks like Twitter? (see informed compliance behaviours)	1= It would seem very bad; 2= It would seem bad; 3= It would seem neither good nor bad; 4= It would seem good; 5= It would seem very good
Legitimacy variables	Moral judgement	Now, thinking in moral terms, how do you think he is performing the following expressions in social networks like Twitter? (see informed compliance behaviours)	If 1= It's totally fine and 5= It's totally wrong
Perceived offensiveness	To what extent do you think someone would be really offended (to the point of it harming their rights) if someone else targeted them with the following type of expression on social networks such as Twitter? (see informed compliance behaviours)		1=Not at all offensive; 5= Totally offensive
Self-censorship variable	Have you ever decided during the last 12 months not to express an opinion on a topic on social networks such as Twitter that you would have liked to express?	1=Never; 2=Some of the times I publish; 3= Quite a few of the times I publish; 4= Almost always or always I publish	

Self-censorship issue	What topics have you avoided expressing your thoughts about at any time during the past 12 months?	1= Never avoided expressing my opinion; 2= Politics; 3= Religion; 3= Economy; 4= Other (describe)
Reasons for self-censorship	In case you ever decided not to express your opinions	1= I have always expressed what I wanted on social networks; 2= I have stopped expressing my opinions because I thought that my message could be removed or blocked; 3= because I thought that my account on the social network could be removed; 4= because I thought that I could be punished for a crime; 5= because I thought that friends, family or close people would not like my opinion; 6= because I thought that others (third parties) would not like my opinion; 7= because I thought that what I wanted to express could offend other people. 8= I believed what I was about to express should not be made public; 9= other