INTERNATIONAL E-JOURNAL OF CRIMINAL SCIENCES
Supported by OMS International Research Centre

"*May I offend you*?" An experimental study on perceived offensiveness in online violent communication and hate speech[1][2]

Rebeca Bautista-Ortuño[3]
Departamento de Psicología de la Salud y Centro CRÍMINA para el estudio y prevención de la delincuencia, Universidad Miguel Hernández de Elche

Francisco J. Castro-Toledo
Centro CRÍMINA para el estudio y prevención de la delincuencia, Universidad Miguel Hernández de Elche

Juan O. Perea-García
Department of Biological Science, National University of Singapore

Nuria Rodríguez-Gómez
Criminologist

Abstract

The aim of this study is to analyze the influence of attitudinal and sociodemographic variables on the perceived offensiveness of online communication and hate speech. We conducted an experimental study in which 373 participants rated their perception of offensiveness of four kinds of violent content (direct incitement to violence/threat, exaltation of violent responses, incitement to discrimination and expression of bad taste) with the appearance of posts in the Facebook social network. We manipulated the emitters (in-group man, in-group woman, and male foreigner). We did find that participants' attitudes towards issues related to the content in the violent messages had a significant effect in how they perceived said content. This kind of investigation becomes extremely

[1] This study has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 740773.
[2] This study has been supported by the National Institute of Cybersecurity (INCIBE) under grant: "*Ayudas para la excelencia de los equipos de investigación avanzada en ciberseguridad*" (reference INCIBEI-2015-02480).
[3] Corresponding author: Rebeca Bautista Ortuño, Avda. de la Universidad s/n, 03202, Elche (España). Tlf. (+34) 966 658 860. e-mail: rbautista@umh.es

INTERNATIONAL E-JOURNAL OF CRIMINAL SCIENCES
Supported by DMS International Research Centre

relevant given the current legal discussion regarding the criminalization of online violent communication and hate speech.

Keywords: *Hate speech, offensiveness, computer-mediated-communication, sociodemographic, attitudinal variables.*

1. Introduction. More than just words: violent communication in social networks

Information and communication technologies (ICT) are increasingly important in our lives, with unwritten norms stemming from the spatio-temporal characteristics of computer mediated communication (CMC, Miró-Llinares, 2012), such as its transnationality, neutrality or lack of censorship for users. The increasing importance of ICT in our lives is made manifest when we look at the massive popularization of the web as a space for interpersonal communication - especially when it comes to social networks such as Facebook or Twitter (Boyd & Ellison, 2007; Rodríguez-Ruibal & Santamaría-Cristino, 2012), with 1500 million users in the case of the former, and 310 million active users (monthly) in the case of the latter, according to official statements from both firms for the year 2016.

Despite the aforementioned particularities of CMC as a medium, it has perpetuated trends of the so-called "violent and hate communication" (Miró-Llinares, 2016) that were already present in face-to-face communication (F2F) or in other previous forms of mediated communication. The latter, traditionally labelled "hate speech", refers to any form of expression directed against groups that are traditionally oppressed on the bases of their identification with a group defined by any combination of race, religion, sexual orientation, functional diversity, ethnicity, nationality, age, gender, social group, political affiliation or, simply, with any group we do not identify with (Miró-Llinares, 2016). Beyond the more classically recognized category of hate speech, researchers have identified another type of violent communication labelled as "words that wound" (Matsuda, Lawrence, Delgado & Williams Crenshaw, 1993). These "words that wound" are verbal expressions that offend or shock a considerable portion of the general public, ultimately triggering penal actions - hence their relevance in relation to the sciences of crime. However, it remains unclear whether all these behaviors can be classified as pertaining to the same category of violent communication. Likewise, much remains to be investigated regarding the endogenous traits that make receivers feel offended by the same mediatic contents, or the emitters out of which they come from (cf., Camps, 2007; Kalinsky, 2004).

2. Literature review

2.1. *Analysis of violent communication and hate speech on the Internet*

Regarding the issue of classifying violent communication, there is a fruitful discussion in the relevant literature, all of them relying on different criteria (cf., Jacks, William & Adler, 2015; Matsuda, 1993; McDevitt, Levin & Bennet, 2002; PROXI, 2015; Sobkowicz & Sobkowicz 2010). However, the emphasis in all of these taxonomies remains on the more fine-grained categorization of hate speech, omitting the aforementioned frame of violent communication, encompassing all speech acts that could be labelled as violent. For the purposes of the present study, we made use of the taxonomy of violent communication and hate speech on the Internet elaborated and justified by Miró-Llinares (2016).

This taxonomy takes a bottom-up approach, by analyzing over 250.000 *tweets* featuring *hate speech* or violent content that were published after the French satirical magazine *Charlie Hebdo* was the target of a terrorist attack. The categorization of the messages into different types allows also for an assessment of their frequency in online communication. The taxonomy was originally conceived as a way to classify the basic categories of violent speech acts while simultaneously assessing the legitimacy of penal action when facing them. It becomes thus easier to question whether criminal law should observe them, or not (cf., Baker, 2007; Boeckmann & Turpin-Petrosino, 2002; Waldron, 2012; Magdy, Darwish & Abokhoadir, 2015, Matsuda, 1993; Miró-Llinares, 2015). It contemplates, first, a categorization based on whether the speech act could result on either a) physical harm, or; b) moral harm, regardless of whether they potentially incur in harming individuals, specific social groups, or the collective of society. The taxonomy, then distinguishes between five categories depending on the emitter's communicative intentions: α) direct incitement to violence/threat of a physical nature; β) exaltation of physical violence; γ) offending honor and/or personal dignity; δ) direct incitement to discrimination; ε) offending collective sensitivity. Lastly, these five categories are divided so as to comprehend all foreseeable modes of violent communication and hate speech, not only in CMC but in any act of human communication (see Table 1 below). We chose Miró-Llinares' (2016) taxonomy because its categories are based, first and foremost, on the analysis of a sample of actual online violent communication in Spanish. Because our participants are also Spanish, and Miró-Llinares' sample was collected recently, we expect a differential perception in offensiveness from our participants depending on the type of message. In other words, we think that the taxonomy is likely to tap on the psychological underpinnings of offensiveness in our sample.

INTERNATIONAL E-JOURNAL OF CRIMINAL SCIENCES
Supported by OMS International Research Centre

Table 1. *Taxonomy of violent communication and hate speech by Miró-Llinares (2016)*

PHYSICAL VIOLENCE	*RELATED TO PHYSICAL HARM*	*α Direct Incitement to Violence*	Direct threat Induction to direct violence Proposition/ Direct provocation to violence
		β Indirect Incitement to Violence	Justify, defend or exalt the execution of violent actions (positive valuation of delinquency) Justify, defend or exalt the execution of violent actions (violence that develops or can develop within a possible legal framework)
MORAL VIOLENCE	*PERSONAL MORAL HARM*	*γ Offending the honor and/or personal dignity*	Honor and/or personal dignity Honor and/or victim's dignity
	CAUSES COMMUNAL MORAL HARM	*δ Discrimination Incitement*	Direct Incitement to discrimination Humiliation, disparagement of one group Expressions of will oriented toward a group Manifestations of defense of past violence
		ε Collective Offense	Expressions of bad taste (gravely impact communal sensibility) Trash talk & especially colloquial expressions in bad taste Expressions of bad taste (though irony) Conspiranoia

However, the focus of this study will not be on which of these behaviors of CMC should be subjected to penal intervention but, rather, to find out the interactions between the

endogenous (who perceives the message) and exogenous (who emits the message) variables that modulate the perception of gravity in potentially offensive messages with the same content. This becomes relevant for the sciences of crime because of the aforementioned propensity of "words that wound" (Matsuda, et al., 1993) to trigger and influence penal actions.

2.2. *Irrelevant demographic variables influence perceived offensiveness*

However, far from unequivocally ascribing a certain quality and degree of harm to violent communication and hate speech, moral judgements are extremely sensitive to biases of both endogenous (who perceives the offensive content) and exogenous nature (who emits the offensive content). The fields of sociology (Tafjel & Turner, 1979) and evolutionary psychology (Cosmides & Tooby, 1992) have identified important cognitive biases in humans to both categorize themselves and others as pertaining to social groups, as well as to behave differently towards other individuals according to whether they are perceived to pertain to the same group (if they are *in-groups*), or to another group (*out-groups*).

The idea of differential treatment of individuals based on physiological, behavioral, psychological, etc. traits that define them as pertaining to different social groups might seem intuitively counterproductive in an increasingly interconnected world. However, there are good reasons to believe that the biases giving rise to these behaviors have been important factors contributing to the survival of our species, with a profound impact in our social interaction still today - indeed humans, like other great ape species (chimpanzees: van Lawick & Goodall, 1968; orangutans: van Schaik, 1999) live in highly dynamic social structures (i.e., *fusion-fission*) that sustain a balance between cooperation and competition among individuals within and across populations (Alford & Hibbing, 2004). Some researchers have hypothesized that it was at least partly the need to keep track of this ever-changing social environment that fostered the development of cognitive abilities in the *homo* genus (Dunbar, 1998). Whichever the overall contribution of these biases to the fitness of our species, it remains clear that apparently irrelevant perceived traits strongly condition the way we conceptualize and behave towards others, challenging the idea that humans act, first and foremost, guided by reason (Cialdini, 2009; Cosmides & Tooby, 1994). As such, we expected participants in our study to vary in their judgement of offensive online content depending on 1) who emitted the message, and; 2) their attitude towards the content of the message.

2.3. *The minimal group effect (MGE) beyond face-to-face interaction*

The most robust evidence we have indicating that we are indeed both extremely prone to establishing social categories and to be sensitive to said categorizations regardless of their real relevance comes from the so called minimal group effect (MGE) (Tajfel, 1970). The MGE denotes the minimal conditions for individuals to show a bias for the groups of which one feels part, or a bias against those groups that one perceives as not belonging to. The paradigm used in this kind of studies aims at forcing participants to take part in a resource distribution task. First, participants are assigned to one of two groups on the basis of arbitrary criteria relative to the task, such as tossing a coin (Rabbie & Horwitz, 1969), sharing preferences about art (Tafjel et al., 1971), or wearing the same color of shirt (Frank & Gilovich, 1988). Then, participants are asked to allocate the limited resources between the two groups. Results tend to show that, although participants are usually fair, they show a significant bias to distribute more of the limited resources to the group they have been allocated to or, conversely, less to the group with which they have not been identified (Brewer, 1979). Therefore, we expected our participants to judge messages as less offensive if they were emitted by an in-group. Conversely, we expected them to be more judgemental when they perceived the emitter to be an out-group.

Amichai-Hamburger (2005) demonstrated the propensity for people to also form groups while surfing the web much like in face-to-face interactions- he divided 24 participants into two different chat groups according to an intuitive choice in a decision-making task. Later, in a cognitive task, group members found the performance of their own group as higher than that of the other group. Due to the affordances of computer mediated communication (CMC), which often consist of impoverished interfaces (at least relative to face-to-face communication, see Miró-Llinares, 2012) researchers have been able to operationalize the MGE in a highly ecologically valid manner in different settings.

Following the trend of studies exploiting the MGE in discrimination, and thanks to the affordances of CMC, Ahmed & Hammarstedt (2008) simulated looking out for an apartment in the Swedish house market by using three different fictional identities - an Arabic/Muslim male, a Swedish female, and a Swedish male. All things standing equal (except the fictional names and e-mail addresses), the Arabic/Muslim male received many less call backs than the Swedish male. It was the Swedish female, however, that met with less difficulty in finding an apartment. Using a very similar paradigm, Von Essen & Karlsson (2013) found that internet auctions also afford MGE on the bases of foreignness and gender. Also using gender as an IV (Alvídrez & Franco-Rodríguez, 2016), show that this perceived demographic trait influences the appraisal that other users make of content

that is, in principle, not relevant to said trait. Building on these studies, we present a MGE paradigm that assesses people's perception of gravity in violent communication and hate speech on the internet, depending on the gender, foreignness, and anonymity of the emitter, while controlling for endogenous variables of the receiver (such as political orientation, gender, religious commitment, etc.). Our hypotheses are 1) that participants will judge messages emitted by out-groups as more offensive than when the same messages are emitted by in-groups, and; 2) that participants' attitude to specific topics will affect the perceived offensiveness of messages dealing with said topics.

3. Methods

3.1. *Sample*

All of our participants (n=373) were Spanish speakers, recruited online through the ad tool provided by the social network *Facebook in* the period comprehended between February 12th and 18th, 2017 (both included), out of which 22.3% were men, and 77.7% were women aged 14-66 (M=23.5; SD=7.9). Of these, 90.9% were single; 8.3% were married, and 0.8% were divorced. Lastly, 43.2% of the participants in our sample received higher education (professional, or graduates), 20.9% were postgraduates, and 2.9% of them received only primary education.

3.2. *Control of endogenous variables*

In order to investigate how the attitudes of our participants influenced their perception of the offensiveness of messages, we controlled for: a) their political orientation - in a Likert scale from 1 ("far left") to 5 ("far right"), we divided participants into three groups with those choosing 1-2 classified as "left wing", those choosing 3 classified as "moderate", and those choosing 4-5 classified as "right wing"); b) their attitude towards bullfighting - in a Likert scale from 1 ("totally against") to 5 ("totally in favor"), again dividing three groups with those choosing 1-2 as being against, those choosing 3 as neutral, and those choosing 4-5 as being in favor; c) religious commitment - with the options "Non-believer", "Agnostic", "Non-practicing believer", and "Practicing believer" we divided participants into two groups depending on whether they chose the two first options (non-believers), or the two last options (believers).

3.3. *Stimuli*

We selected four messages, each of them representing categories identified by Miró-Llinares in his taxonomy (2016). Message α: Direct incitement to violence/threat. Content: [an image shows the advertisement for a well-known Spanish stand up comedian]. Emitter's message: "Next Saturday there's this clown coming to the theater, would someone sign up to give him a good scare? It should suffice if you just bring some sticks and knives. Message β: exaltation of violent responses. Content: [an image shows a scene of a popular Spanish celebration particular to the town of Tordesillas, consisting on physically harassing a bull to death with the headline *The City Council of Tordesillas, sentenced for the illegal celebration of the* Toro de la Vega *in 2014*"]. Emitter's message "And so to jail with a few of these wretches!!! The town deserves getting bombed… they don't deserve to live". Message δ: incitement to discrimination. Content: [Sport news on basketball with the headline "*The Maccabi beats Real Madrid by 98-86 in the Euro finals*", with an image where a basketball player from Real Madrid looks visibly disappointed]. Emitter's message: "Fucking jews, Hitler should have exterminated all of them in the gas chamber. Damn bastards!". Message ε: expression of bad taste. Content [an image shows former Pope Ratzinger approaching a baby with the headline "Undignifying! The Pope says that child abuse is not that bad, that it was commonplace back in his days…"]. Emitter's message: "Thanks, but I don't like them so young…".

INTERNATIONAL E-JOURNAL OF CRIMINAL SCIENCES
Supported by OMS International Research Centre

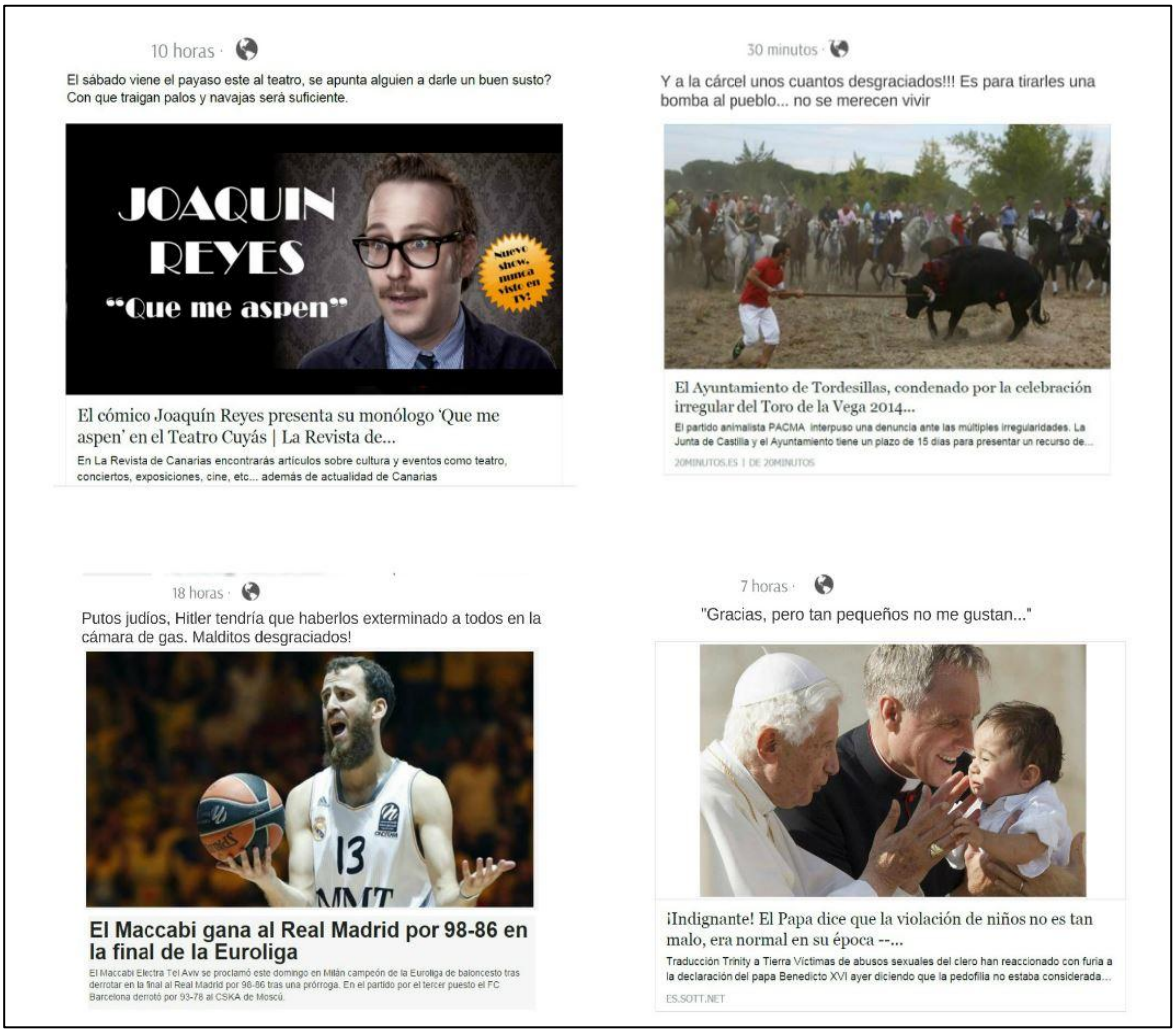

Figure 1. Selected messages as stimuli

Our stimuli were designed so that they resembled those published in the *Facebook* social network. These four messages were, in turn, combined with four different fictional emitters, marked by their displayed names in the typical *Facebook* format - a national male in-group or MI-G ("Antonio López"); a national female in-group or FI-G ("Laura Gonz."), a person of islamic or arabic ascent, representing a male out-group or O-G ("Abd Al-Hamid U.") and, lastly, a control emitter, featuring a pixelated name, so that it was indistinguishable). We chose to not include a picture of our make-do users of *Facebook*

so as to avoid effects derived from random confounding variables such as attractiveness, etc. The original stimuli (in Spanish) can be seen in Appendix I.

3.4. *Procedure*

An experimental design with 16 conditions resulting from the combination of the four different messages with the four different emitters. Participants, who were randomly assigned one out of the 16 conditions, had then to assess the perceived offensiveness of the four messages in the format that is characteristic of Facebook. The degree of perceived offensiveness was measured with the following item: "In your opinion, how offensive is [AVATAR`S NAME] ´s comment about the news he/she has shared?". Responses were on a 1-5 Likert scale (1=not offensive at all, 5=completely offensive). Lastly, participants had to fill in the details regarding their own attitude towards several current issues and demographics (political orientation, religious commitment, gender, etc.). The survey was distributed through paid marketing services in *Facebook*. We used Google's free survey system to elaborate the survey. The criteria to be eligible to take part in the study were: 1) being a Facebook user and living in Spain; 2) being at least 13 years old, and; 3) speaking Spanish. The campaign ran for a week, obtaining a total of 373 valid responses.

3.5. *Analysis*

We used the statistical package IBM Statistics SPSS v. 24 for the quantitative analysis of the data. We ran a between-subjects one-way ANOVA to analyze: 1) the differences in the perceived offensiveness of each message, as a pretest, and 2) the influence in the evaluation of the offensiveness of each message of a) the type of emitter, b) the political orientation of the receiver, c) the simple effects of the interaction between the variables Religious feeling and type of emitter in the message ε (religion related content), and d) the simple effects of the interaction between the variables Attitude toward the bullfighter and type of sender in the message β (of bullfighting related content). Finally, to analyze the influence of receiver gender on the perception of offensiveness of the messages according to each type of emitter, we applied an independent measures Student T-test. We considered results to be significant if they reached a value equal or smaller than 0.05. Lastly, we also calculated estimations of the size of the effect for each of the tests we ran.

4. Results

We first run an ANOVA to test the means in the perceived offensiveness in our control condition (ie., "anonymous facebook user"). The results (see table 2) show significant differences in the perceived offensiveness in each of the messages.

Table 2
Comparison of the means for perceived offensiveness in each of the messages in the control condition (ie., "anonymous facebook user").

Message	N	Average	SD	F	*p*	Effect size
α	79	4.32	1.03	34.710	0.000	0.22
β	98	3.67	1.24			
δ	106	4.80	0.51			
ε	90	3.41	1.31			

We run a Sheffe test to find out between which of the messages there were significant differences. This turned out to be true between all of our messages except β and ε (difference between the means=0.26; p=0.324). As such, it can be concluded that the most offensive message by our participants would be δ (Discrimination incitement). The next most offensive one would be α (Direct incitement to violence), followed by β and ε, respectively (Indirect incitement to violence and collective offense).
In general, perceived offensiveness was high, with the lowest scores found for message β (exaltation of violent responses), and message ε (expression of bad taste). Table 3 below summarizes these results. Table 3 also summarizes the perception of offensiveness according to the emitter - as it can be observed, emitters do not seem to affect the perception of offensiveness of each of the messages in a significant way. We can therefore state that our hypothesis that participants will judge messages emitted by out-groups as more offensive than the same messages emitted by in-groups is not confirmed.

INTERNATIONAL E-JOURNAL OF CRIMINAL SCIENCES
Supported by OMS International Research Centre

Table 3
Comparison of the means for perceived offensiveness in each of the messages, according to the emitter.

Message	Emitter	N	Average	SD	F	*p*	Effect size
α Direct incitement to violence	MI-G	88	4.24	0.97	0.102	0.959	0.001
	FI-G	106	4.28	0.88			
	O-G	100	4.29	0.86			
	Control	79	4.32	1.03			
β Indirect incitement to violence	MI-G	88	3.64	1.37	1.102	0.348	0.009
	FI-G	85	3.67	1.31			
	O-G	102	3.92	1.06			
	Control	98	3.67	1.24			
δ Discrimination incitement	MI-G	94	4.81	0.47	1.293	0.276	0.01
	FI-G	85	4.67	0.70			
	O-G	88	4.72	0.59			
	Control	106	4.80	0.51			
ε Collective offense	MI-G	103	3.47	1.46	0.295	0.829	0.002
	FI-G	97	3.42	1.45			
	O-G	83	3.59	1.36			
	Control	90	3.41	1.31			

An analysis of the ratings according to the demographics of our participants did show significant results for several messages. Independently of the emitter, participants' gender played a role in assessing their offensiveness. Women ranked message message α (direct incitement to violence/threat) and message δ (incitement to discrimination) as significantly more offensive than men. These results are summarized in table 4 below.

Table 4
Comparison of the means of perceived offensiveness according to the receiver's gender.

Message	Sex	N	Average	SD	T	*p*	Effect size
α Direct Incitement to Violence	Man	83	4.01	1.16	-3.037	0.003**	0.16
	Woman	290	4.36	0.83			
β Indirect Incitement to Violence	Man	83	3.78	1.24	0.424	0.671	0,.02
	Woman	290	3.72	1.25			
δ Discrimination Incitement	Man	83	4.61	0.70	-2.174	0.032*	0.11
	Woman	290	4.79	0.52			
ε Collective Offense	Man	83	3.39	1.40	-0.618	0.537	0.03
	Woman	290	3.49	1.40			

*significance at 0.05; **significance at 0.01

Our participants' political orientation did not influence their responses in a significant manner - except for message δ (incitement to discrimination), participants identifying themselves as being of a left-wing ideology tended to perceive each of the messages as less offensive than those identifying themselves with a right-wing ideology. However, the ANOVA returned no significant differences depending on the political orientation of our participants. The results are summarized in table 5 below.

INTERNATIONAL E-JOURNAL OF CRIMINAL SCIENCES
Supported by OMS International Research Centre

Table 5

Comparison of the means of perceived offensiveness depending on the receiver's political orientation.

Message	Political orientation	N	Average	SD	F	*p*	Effect size
α Direct Incitement to Violence	Left-wing	165	4.25	0.93	0.311	0.733	0.002
	Moderate	166	4.28	0.98			
	Right-wing	42	4.38	0.66			
β Indirect Incitement to Violence	Left-wing	165	3.56	1.27	2.856	0.059	0.015
	Moderate	166	3.84	1.21			
	Right-wing	42	3.95	1.25			
δ Discrimination Incitement	Left-wing	165	4.75	0.59	0.024	0.976	0.000
	Moderate	166	4.76	0.56			
	Right-wing	42	4.74	0.50			
ε Collective Offense	Left-wing	165	3.28	1.44	2.810	0.062	0.015
	Moderate	166	3.58	1.35			
	Right-wing	42	3.74	1.33			

Unsurprisingly, participants who considered themselves more religiously committed (ie., "believers") considered message ε (bad taste) as more offensive than those who considered themselves non-believers, independently of who was making the statement (see table 6 below). A Student T-test also bore significant differences when the message is emitted by a foreigner, or an anonymous person

INTERNATIONAL E-JOURNAL OF CRIMINAL SCIENCES
Supported by OMS International Research Centre

Table 6
Comparison of the means of perceived offensiveness of "bad taste" depending on the emitter and the religious commitment of the receiver.

Message	Religious commitment	N	Average	SD	F	*p*	Effect size
ε Collective Offense /Man	Non-believer	69	3.42	1.51	0.203	0.653	0.002
	Believer	34	3.56	1.38			
ε Collective Offense /Woman	Non-believer	67	3.33	1.44	0.917	0.341	0.010
	Believer	30	3.63	1.47			
ε Collective Offense /Foreigner	Non-believer	49	3.33	1.41	4.692	0.033*	0.055
	Believer	34	3.97	1.22			
ε Collective Offense /Anonymous	Non-believer	52	3.04	1.31	11.174	0.001**	0.113
	Believer	38	3.92	1.12			
		N=373					

*significance at 0.05; **significance at 0.001

Because message β (exaltation of violent responses) was operationalized with a bull-fighting theme, it is also not surprising that participants with a favorable attitude towards bullfighting found its offensiveness greater than those who did not consider themselves favorable to bullfighting (see table 7 below). However, we found significant differences only when the emitter was a man. These results, together with those in table 6, give support to our second hypothesis, that participants' attitude to specific topics will affect the perceived offensiveness of messages dealing with said topics.

INTERNATIONAL E-JOURNAL OF CRIMINAL SCIENCES
Supported by OMS International Research Centre

Table 7
Comparison of the means of perceived offensiveness of message β (exaltation of violent responses) according to both kind of emitter, and receivers' attitude towards bullfighting.

Message	Attitude	n	Average	DT	F	*p*	Effect size
β Indirect Incitement to Violence /Man	Against	60	3.42	1.37	4.289	0.017*	0.092
	Neutral	19	3.79	1.44			
	For	9	4.78	0.44			
β Indirect Incitement to Violence /Woman	Against	64	3.56	1.34	1.102	0.337	0.026
	Neutral	11	4.18	1.08			
	For	10	3.80	1.32			
β Indirect Incitement to Violence /Foreigner	Against	83	3.88	1.10	0.507	0.604	0.010
	Neutral	12	4.00	0.95			
	For	7	4.29	0.76			
β Indirect Incitement to Violence /Anonymous	Against	71	3.55	1.32	2.572	0.082	0.051
	Neutral	14	3.64	0.84			
	For	13	4.38	0.96			
		N=373					

*significance at 0.05

5. Discussion and Conclusions

In the present study, we manipulated the perceived emitter of four fictitious posts with the same format as the *Facebook* social network with contents that were classified as

either *violent*, or *hate speech*. Despite the negative results testing our first hypothesis (see table 3), previous research has investigated the influence of the emitters' demographic factors in the perception of the same content - for example, Alvídrez & Franco-Rodríguez (2016) manipulate the emitter's gender in Twitter messages, finding that both style (sudden or direct vs. submissive or indirect) and gender (male or female) were heavily associated with the level of credibility and persuasion, as measured by the capacity to attract and involve other users in public events. Similarly, Von Essen & Karlsson (2013) showed that the perceived gender and foreignness of sellers in online auctions affected buyer discrimination. Also, Ahmed & Hammarstedt's (2008) results support the idea that irrelevant demographic factors such as perceived foreignness and gender have an effect in the degree of success of their online personas in finding housing in the online Swedish house market. Our results replicated these findings only partially, since we found that participants who perceived β as significantly more offensive did so only when the emitter was a man.

Regarding our second hypothesis, our results show that these endogenous factors influence the degree of perceived offensiveness in online *hate speech*, or in online messages with violent content, in line with previous research pointing at a strong link between self-identification with a group, and intra-group favoritism (Brown et al., 1986; Brown & Williams, 1984; Dasgupta, 2004; Hinkle & Brown, 1990). For example, our participants that admitted to have a favorable view of bullfighting considered message β as significantly more offensive than other participants. Similarly, participants who considered themselves religious ranked message ε as more offensive than other participants.

Our results are interesting despite showing no statistically significant influence of the independent variable (ie., emitter), as can be seen in table 3 above. However, there is a clear pattern in the way our participants assessed the different messages entirely based on their content, if we contrast the assessments with the categories in Miró-Llinares (2016). Namely, perceived offensiveness stands in an inverse relation with the frequency of each of the categories. That is, the more our participants rated the message as offensive, the least frequent the category that it belongs to. This is easily observable if we group categories α and β (causing physical harm) on the one hand, and δ and ε (causing moral harm) on the other (see table 1). The message belonging to category α was rated as the most offensive and stands as the least frequent (0.1%). Category β, the second most offensive, is the second least frequent (3.7%). Categories δ and ε, rated the least offensive are, conversely, the most common types (26% and 67.8%, respectively).

We think these results are interesting in light of current political-criminal discussions in our legal ecosystem, with many scholars proposing that potentially offensive communication should be, or not, contemplated by criminal law (cf., Baker, 2007; Boeckmann y Turpin-Petrosino, 2002; Waldron, 2012; Magdy, Darwish y Abokhoadir, 2015; Matsuda, 1993; Miró-Llinares, 2015, among others). Our results, together with the aforementioned research, strongly suggest that we revisit our assumptions regarding the factors that influence punitive responses in both laymen and experts closely linked to legal and judicial decision-making processes. This is because, far from responding to criteria that are inherent to the formal aspects of a message, these responses are significantly influenced both by the emitters' demographic factors, as well as by the attitude, world-view etc., of the receiver. The danger lays in that, in principle, these factors could be intrinsically irrelevant to the assessment in question. We think this calls for an incorporation of contextual factors in the investigation of the effects of violent communication and *hate speech* (Cowan & Hodge, 1996) in CMC.

Lastly, our study, while less ecologically valid (since it measures explicit self-report data, rather than implicit behavioral outcomes), granted us the possibility to control for data regarding our participants' world-views, attitudes towards specific topics, and demographic data. As such, our paradigm offers the chance to explore the interaction between the emitter's perceived identity, and variables that are endogenous to our participants, such as their attitude towards bullfighting, religion, or politics. On the other hand, we think that some current limitations of our paradigm are worth noting for future research - firstly, we believe that social desirability might have played a role in the responses we received from our participants by enhancing out-group favoritism (Evans et al., 2003; Lynch & Addintong, 2010; Tynes, Reynolds, & Greenfield, 2004), effectively lowering their score of perceived offensiveness for our stimuli. Likewise, our Likert scale on five points might have been too coarse grained to obtain significant differences from the different perceived emitters. Another issue that has to do with our methodology is that we chose not to explicitly indicate age, gender, ethnicity, etc. of our make-do *Facebook* users. Whereas this increases the ecological validity of our design, it also means that we cannot be sure that participants paid attention to the relevant variables that indicated sex, or foreignness (ie., the names). This might have been further accrued by the choice of not including graphic representations of our make-do *Facebook* users (ie., *profile pictures*). Our proposal for further research using this paradigm would be therefore to collect behavioral data by complementing it with, for example, *eye-tracking*, which could be useful in indicating whether our participants paid attention to the relevant variables, as well as discerning patterns that might arise from self-aware inhibitory control when scoring out-groups.

References

Ahmed, A. M., & Hammarstedt, M. (2008). Discrimination in the rental housing market: A field experiment on the Internet. *Journal of Urban Economics, 64*(2), 362-372.

Alford, J. R., & Hibbing, J. R. (2004). The origin of politics: An evolutionary theory of political behavior. *Perspectives on Politics, 2*(04), 707-723.

Alvídrez, S. y Franco-Rodríguez, O. (2016). Powerful Communication Style on Twitter: Effects on Credibility and Civic Participation. *Comunicar, 47*(24), 89-97.

Amichai-Hamburger, Y. (2005). Internet minimal group paradigm. *CyberPsychology & Behavior, 8*(2), 140-142.

Baker, D. J. (2007) The Moral Limits of Criminalizing Remote Harms Review. *New Criminal Law Review: An International and Interdisciplinary Journal, 10*(3). 370-391.

Boeckmann, R. J.; Turpin & Petrosino, C. (2002). Understanding the harm of hate crime. *Journal of Social Issues*, *58*(2).

Boyd, D. M. & Ellison, N. B. (2007). Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication*, *13*, 210-230.

Brewer, M. B. (1979). In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. *Psychological bulletin, 86*(2), 307.

Brown, R., & Williams, J. (1984). Group identification: The same thing to all people?. *Human Relations, 37*(7), 547-564.

Brown, R., Condor, S., Mathews, A., Wade, G., & Williams, J. (1986). Explaining intergroup differentiation in an industrial organization. *Journal of Occupational psychology, 59*(4), 273-286.

Camps, V. (2007). Ofensas y Libertad de expresión. *QUA-DERNS*, *27,* 3-12

Cialdini, R. (2009). *Influence: Science and Practice*. Boston, MA: Pearson Education.

Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. In J, Barkow, L. Cosmides & J. Tooby (eds.), *The adapted mind: evolutionary psychology and the generation of culture* (pp. 163-228). New York: Oxford University Press

Cosmides, L., & Tooby, J. (1994). Better than rational: Evolutionary psychology and the invisible hand. *The American Economic Review, 84*(2), 327-332.

Cowan, G., & Hodge, C. (1996). Judgments of hate speech: The effects of target group, publicness, and behavioral responses of the target. *Journal of Applied Social Psychology*, *26*(4), 355-374.

Dasgupta, N. (2004). Implicit ingroup favoritism, outgroup favoritism, and their behavioral manifestations. *Social Justice Research*, *17*(2), 143-169.

Dunbar, R. I. (1998). The social brain hypothesis. *Brain*, *9(*10), 178-190.

Evans D. C., Garcia D. J., Garcia D. M. & Baron R. S. (2003). In the privacy of their own homes: Using the Internet to assess racial bias. *Personality and Social Psychology Bulletin 29*(2), 273-284.

Frank, M. G. & Gilovich, T. (1988). The dark side of self and social perception: Black uniforms and aggression in professional sports. *Journal of Personality and Social Psychology, 54,* 74-83

Hinkle, S., & Brown, R. (1990). Intergroup comparisons and social identity: Some links and lacunae. *Social identity theory: Constructive and critical advances*, *48*, 70.

Hughes, M. G., Griffith, J. A., Zeni, T. A., Arsenault, M. L., Cooper, O. D., Johnson, G., Hardy, J. H., Connelly, S. and Mumford, M. D. (2014), Discrediting in a Message Board Forum: The Effects of Social Support and Attacks on Expertise and Trustworthiness. *J Comput-Mediat Comm, 19*, 325–341.

Jacks, W.; Adler, J. R. (2015). A proposed typology of online hate crime. *Psychology Unbound*: *Open Access Journal of Forensic Psychology*,*7,* 64-89.

Kalinsky, B. (2004). El contexto de la ofensa: un concepto significativo para el análisis del delito. *Urbe et ius: revista de opinión jurídica*, *2*, 160-175.

Lynch, J. P. & Addintong, L. A. (2010). Identifying and Addressing Response Errors in Self-Report Surveys. En A. R. Piquero & D. Weisburd (eds.), *Handbook of Quantitative Criminology* (pp. 251-272). New York: Springer

Magdy, W., Darwish, K. & Abokhoadir, N. (2015). *Quantifying Public Response towards Islam on Twitter after Paris Attacks.* arXiv preprint arXiv:1512.04570.

Matsuda, M. J., Lawrence III, C. R, Delgado R., & Williams Crenshaw, K. (1993). *Words that wound. Critical Race Theory, assaultive speech and the First Amendment*. Oxford: Westview Press.

Matsuda, M. J. (1993). *Words that wound: Critical race theory, assaultive speech, and the first amendment*. Westview Press.

McDevitt, J., Levin, J. & Bennett, S. (2002). Hate Crime Offenders: An Expanded Typology. *Journal of Social Issues*, *58*(2), 303-317

Miró-Llinares, F. (2012). *El cibercrimen. Fenomenología y criminología de la delincuencia en el ciberespacio.* Madrid: Marcial Pons.

Miró-Llinares, F. (2015). La criminalización de conductas "ofensivas": A propósito del debate anglosajón sobre los" límites morales" del derecho penal. *Revista electrónica de ciencia penal y criminología*, *17*. 1-65.

Miró-Llinares, F. (2016). Taxonomía de la comunicación violenta y el odio en Internet. *Revista de Internet, Derecho y Política*, *22,* 82-107.

PROXI (2015). Report from Observatorio "Proyecto Online contra la Xenofobia y la Intolerancia" (Vol I). Retrieved from http://www.observatorioproxi.org/im'ages/pdfs/INFORME-proxi-2015.pdf.

Rabbie, J. M., & Horwitz, M. (1969). Arousal of ingroup-outgroup bias by a chance win or loss. *Journal of personality and social psychology, 13*(3), 269.

Rodríguez-Ruibal, A., & Santamaría-Cristino, P. (2012). Análisis del uso de las redes sociales en Internet: Facebook y Twitter en las universidades españolas. *Revista de comunicación y tecnologías emergentes, 10* (2), 228-246.

Sobkowicz, P., & Sobkowicz, A. (2010). Dynamics of Hate Based Internet User Networks. *The European Physical Journal B*, *73*(4), 633-643.

Tajfel, H. (1970). Experiments in intergroup discrimination. *Scientific American*, *223*, 96-102

Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. *The social psychology of intergroup relations, 33(*47), 74.

Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behaviour. *European journal of social psychology, 1*(2), 149-178.

Tynes, B., Reynolds, L., & Greenfield, P. M. (2004). Adolescence, race, and ethnicity on the Internet: A comparison of discourse in monitored vs. unmonitored chat rooms. *Journal of Applied Developmental Psychology*, *25*(6), 667-684.

Van Lawick-Goodall, Jane (1968). The Behaviour of Free-Living Chimpanzees in the Gombe Stream Reserve. *Animal Behaviour Monographs (Rutgers University*), 1(3), 167

Von Essen, E., & Karlsson, J. (2013). A matter of transient anonymity: Discrimination by gender and foreignness in online auctions. Available at SSRN 2154848.

Waldron, J. (2012). *The harm in hate speech*. Harvard University Press.