VOX POL WORKSHOP: ETHICS AND POLITICS OF ONLINE MONITORING OF VIOLENT EXTREMISM

Working paper[*] Presented at VoxPol Workshop,
Brussels, 19th – 20th January, 2015

Monitoring racist and xenophobic extremism to counter hate speech online: ethical dilemmas and methods of a preventive approach

Andrea Cerase[1]*, Elena D'Angelo*[2]*, Claudia Santoro*[3]

Extended abstract

In recent years online racism has seen a quick and serious growth in many European and non - European countries, till to become a worrying global phenomenon (Perry & Olsson, 2009). One of the most striking examples of such process is the rise of White Supremacist Movements online, whose strategy mainly consists in disguising their hidden political agenda and attempting to subvert civil rights by presenting their standpoints through an overturn of the rhetoric of the civil rights movement, aimed to destroy civil rights themselves (Daniels, 2009). Undoubtedly, such kind of hidden racist expressions are intended to exploit "favorable" conditions as the financial crisis, the increase of social conflict and the rise of populist issues in politics. In Italy, as an example, UNAR, a governmental anti-discrimination body, documented that complaints for online racism weighed for 30.9% of the overall cases involving the media (UNAR, 2013). Similar situations have been also found in other European countries such as Slovenia, Finland, Hungary and United Kingdom, as it emerged by part of the research carried out within the European project LIGHT ON (Boileau, Del Bianco, Velea, 2014)[4]. Nowadays, racist claims are not only pursued in terms of blatant racism. In fact, a huge amount of disguised racist contents is currently published on the Internet in form of

[*] PLEASE NOTE: this *working paper* should be intended as a preliminary draft: please do not cite or circulate without the authors' permission.

[1] Sociologist, at Regione Abruzzo

[2] Project Officer at UNICRI, Turin

[3] Project Officer at Progetti Sociali, Pescara

[4] LIGHT ON project, JUST/2012/FRAC/AG/2699, co-financed by Fundamental Rights and Citizenship of the European Commission.

occasional bigotry or individuals' outburst, whereas they are, as a matter of fact, intended to foster racist attitudes among people and to support racism normalization. Some scholars defined as "common sense racism" or "rational racism" the attempt to talk against immigrants, refugees, minority members (as well as homosexuals and disabled) as undesirable, avoiding to be labeled as "racist" (Capdevila and Callaghan, 2008; Meddaugh and Kay, 2009).

The Internet is playing a crucial role in the so-called normalization of racism: racist movements are well aware about the potential of social media in the diffusion of hate speech, and they exploit virality by cloaking the real source(s) of such messages to manipulate people worries and outrage and to promote sharing of such contents (Andrisani, 2014). Some scholar argued: "One gets the impression that they are now collapsed the weak shelters of censorship and self-censorship that until recent years made difficult to pronounce explicitly racist discourses in public: topics such as «anti-social behaviors» of the Gypsies, the identification of immigration with crime, the danger of certain «races», the invitation to sink the boats of the immigrants - that once a time was a preserve of the [Northern] League's[5] rhetoric [in Italy] - are now spoken publicly without any shame, and not only by right wing speakers" (Rivera, 2008). This complex and controversial issue refers to an important ethical dilemma: *how can we tackle this phenomenon without undermining freedom of speech?*

The question has been effectively summarized by Nils Muižnieks, Commissioner for Human Rights at the Council of Europe. As it is recognized, *hate speech* is not about *freedom of speech*: it's a threat against the rights of others and public safety, since hate speech and violent action appear to be tightly intertwined. Many incidents occurred in the recent past can easily demonstrate how hate speech can be actually perceived as an authorisation to engage in violence, which is likely to lead in committing real-life crimes, therefore it is necessary to deal simultaneously with hate speech and hate crimes (Muižnieks, 2013).

Populist movements often exploit arguments likely to be racist in their discourse; in all cases allegations of racism, and especially of racist hate speech, must be substantiated by evidences and every complaint must be proven to be realistic. Since "hate speech" is per se a controversial concept, and at the moment broadly accepted definitions are still missing, policy makers and law enforcement agencies must deal with intrinsic ambiguity and polysemy of such contents. Given that these messages can result in different interpretation by different people, it is more essential to ground reporting, assessments and legal prosecution on objective and factual arguments, taking carefully into account all the available information about the source and the context in which messages have been spread. Indeed, on line communication strategy of hate mongers is often grounded on old style propaganda techniques, in which authorship, source, or real intention of a publication or broadcast are intentionally cloaked or disguised (Daniels, 2009: 119)

[5] The *Lega Nord* (Northern League) is a federalist and regionalist political party in Italy, established in 1991 by Umberto Bossi. This party advocates for secession of the North of Italy and its members are very often involved in racist and xenophobic political talk (see Avanza, 2010).

The transnational experience of the LIGHT ON project shows that defining whether a content is racist or not can be facilitated by joining experiences from different sectors. Indeed, investigations on a suspected hate speech case, and the related prevention policies, must take into account diverse information about the source and the context in which messages have been spread: “collecting and analyzing the different expressions of contemporary racism is essential to understand the phenomenon and to design new strategies contrasting it” (Boileau, Del Bianco, Velea, 2014). Furthermore, looking at this type of online monitoring from a broader perspective, it emerges clearly that a transnational approach is needed, as it is demonstrated by the findings of the research carried out by national teams in five European countries, which are presented on the LIGHT ON website using tools such as a visual database and a collaborative glossary. These digital tools, that allow users to filter entries per country, target, typology and target group, demonstrate “how European Nazi and Fascist groups are tightly connected, in order to create a wide racist network across Europe. Many racist watchwords and symbols are indeed the same in different countries and Nazi websites often have many shared inbound and outbound links with their correspondents in other countries” (Cerase, 2014). In particular, the LIGHT ON project research has found many examples, ranging from the so-called “black” propaganda, aimed to deceive target groups by spreading false material through a disguised source, to the “grey” propaganda, in which sources are not identified and contents are at least partially true, although carefully selected to induce certain effects such as persuasion and mobilization (see Jowett & O’Donnel, 1992; Mc Quail, 2000).

Some of the main tools developed within the project LIGHT ON (database, glossary, training manual, toolkit, guide on 'how to spot online racism' etc.) can provide some non-arbitrary evaluation parameters in order to differentiate between what constitutes an actual instigation to hate from what does not, also in view of the ethical dilemma concerning hate speech vs freedom of speech (further discussed later on in this paper) – unfortunately at the core of recent dramatic events happened in Paris.

One of the main aims of the LIGHT ON research was to investigate modern verbal and visual manifestations of racism and xenophobia. Very often racist symbols and images are accepted as normal social expressions, but as they convey much meaning, intent and significance in a communicative and immediately recognizable form, they influence personal and collective behaviors, especially when they are shared on the Internet, where these visual expressions can easily engage broad audiences. “These ‘newer’ forms of racism are so embedded in social processes and structures that they are even more difficult to explore and challenge” (Bajt, 2014). These expressions are used as tools to carry out racist arguments, raise the level of violence tolerated by the society and lead to a dangerous normalization of racism. Indeed, the European Commission against Racism and Intolerance of the Council of Europe warned how “such public manifestations risk fuelling racism, xenophobia, anti-Semitism and intolerance” (ECRI, 2013).

The LIGHT ON project brought together experts working in different fields, bringing their contribution on the state of national legislations, political issues on national agendas, policies for prevention, media reporting and relevant academic literature. They contributed to the data collection phase with their multi-sectorial expertise: victim support groups, for example, stressed how certain contents have the precise scope to harm people, and thus must be discerned from jokes, irony and satire, and labeled for their strong negative impact. Such synergy among different stakeholders suggested the importance of a victim-centered approach, and also the importance of the cooperation among civil society, researchers, local authorities and law enforcement agencies. It also stressed the role of national authorities and groups providing legal service support in following up racist cases emerged from the monitoring or reported by victims. Indeed, despite online monitoring can be intended as a deterrent and a tool to prevent - or control – racist contents, its effectiveness results stronger when monitoring is linked to the action of support groups or national authorities at national level, not only for ethical reasons, but also to encourage self-reporting.

Going back to both prevention and action against hate speech, among the most important findings of the LIGHT ON project, the monitoring role of users and their key function in reporting online hate speech and racist propaganda emerge as essential tools in the fight against these phenomena, also in view of improving and increasing the response of the law enforcement authorities. Understanding the main reasons why hate speech on the Net often go unreported is a starting point to draw guidelines for reporting and tips for monitoring online materials promoting violent extremism (FRA, 2012). Some of these reasons are linked to the lack of confidence in the police by the victims, concern about revenge attacks or fear of retaliation, acceptance of violence and abuse (*nothing will change anyway!*), fear of having privacy compromised, fear of jeopardising immigration status, cultural language barriers or lack of victim support system.

Moving in this perspective the LIGHT ON project elaborated, within the training manual on *Investigating and Reporting Hate Speech Online* (UNICRI, 2014), a set of general tips for online reporting, with a particular focus on the main social networks as one of the main vehicles for spreading violent extremism and populist propaganda.

One of the first “tips” for users to report correctly an online hate incident is to evaluate the content of the speech and select the best strategy accordingly (Mnet, 2012). The user should consider whether the content is hosted in his/her own country and thus subjected to national legislation. However authors are well aware that this could make their identification easier, therefore they place the contents violating the national rules on servers located abroad. The main suggestion in this case for the users is: always have a backup of the content of the hate speech incident! (and the LIGHT ON training manual includes a list of different concrete steps on how to backup, among other tips).

The main steps for reporting violent extremism on the most used social media (*Facebook, Twitter, Wikipedia* and *You Tube*) are also outlined in the LIGHT ON training materials: acquiring this knowledge makes it easier for law enforcement to adopt a victim-centered approach and effectively help victims by pointing them to the right path of reporting online. The different social networking sites have different policies on definition of hate speech and methods to report and/or block the contents. Moreover, even when the online reporting fails, ISPs and Social Networking companies may have established policies to collaborate more efficiently with national authorities on the regulation and removal of hate speech.

Going back to the importance of recognizing hatred contents, in order to correctly treat them, the huge dilemma regarding the relation between populist arguments and violent extremism on the Internet, on one side, and freedom of speech on the other can not be left aside, especially in these dramatic days in Europe (CoE, 2012).

"Reconciling rights which are at the core of democracy, such as freedom of belief and religion and freedom from discrimination, with the right to freedom of expression represents a significant challenge. When comedy and dark humour are included in the picture, establishing clear boundaries between what constitutes freedom of expression and what falls under the category of hate speech becomes an ever more complex challenge" (UNICRI, 2014). But where do we draw the line?

Even if comedy and satire, as forms of expressions, are protected by laws dealing with freedom (of expression) they also come with duties and responsibilities and, as such, may be subjected to restrictions or penalties as prescribed by law. This implies that in democratic societies, governments may limit freedom of expression where necessary, but only in so far as they are provided for by law and in a manner which is proportionate. The test against which such limitations are evaluated is a strict one.

References

– Andrisani, P. (2014) Quando il razzismo nel web diventa “virale” in Centro Studi e Ricerche IDOS (eds.) *Rapporto Unar. Dalle Discriminazioni ai diritti*, IDOS, Roma, pp. 249-252

– Avanza, M. (2010). The Northern League and its ‘innocuous’ xenophobia. In Mammone, A., & Veltri, G. A. (Eds.). (2010). *Italy today: The sick man of Europe*. Routledge, Abingdon, UK, 131 – 142

– Bajt, V., (2014). “Contemporary racism across Europe”, *Freedom From Fear Magazine*, 9, 36-41.

– Boileau, A., Del Bianco D., Velea, R. (eds., 2014), *Understanding the perception of racism. Research as a tool against racism*, Light On Project, Gorizia (ISBN 978-88-89825-32-7)

– Capdevila, R., and Callaghan, J. E. (2008). ‘It's not racist. It's common sense’. A critical analysis of political discourse around asylum and immigration in the UK. *Journal of Community and Applied Social Psychology*, *18*(1), 1-16.

– Cerase, A. (2014) “Racist symbols and discourses: from Essentialist to Far-right racism”, *ENARgy The European Network Against Racism's webzine*, April 2014, http://www.enargywebzine.eu/spip.php?article349

– CoE, (2012), “Cyberhate and freedom of expression” (paragraph 3), in *Mapping study on projects against online hate speech.* DDCP-YD/CHS (2012) 2, Coe, Strasbourg.

– Daniels, J. (2009). *Cyber racism: White supremacy online and the new attack on civil rights*. Rowman & Littlefield Publisher

– ECRI (2013) *Annual Report on Ecri’s Activities*, (1 Jan . 31 Dec, 2012) Council of Europe, Strasbourg.

– FRA, 2012, *Making hate crime visible in the European Union: acknowledging victims' rights*, FRA – European Union Agency for Fundamental Rights, Bruxelles

– Jowett, G. S. O’Donnell V., (1992) *Propaganda and Persuasion*. London: Sage.

– McQuail, D. (2000). *Mass media theory: An introduction*. London: Sage.

– Meddaugh, P. M., and Kay, J. (2009). Hate Speech or “Reasonable Racism?” The Other in Stormfront. *Journal of Mass Media Ethics*, *24*(4), 251-268.

– Muižnieks, N. (2013) Hate speech is not protected speech *ENARgy The European Network Against Racism's webzine* http://www.enargywebzine.eu/spip.php?article332

– Mnet, 2012, *Responding to Online Hate, Media Awareness* Network,Ottawa, Canada

– Perry, B., & Olsson, P. (2009). Cyberhate: the globalization of hate. *Information & Communications Technology Law*, *18*(2), 185-199.

– Rivera, A.M (2008) La *normalizzazione* del *razzismo* in Naletto G. (ed.), *Sicurezza di chi? Come combattere il razzismo*, edizioni dell'asino, Roma: 55-61

– UNAR (2013) *Relazione al Presidente del Consiglio dei Ministri sull’attività svolta dall’UNAR – Ufficio per la promozione della Parità di Trattamento e la Rimozione delle Discriminazioni Fondate sulla Razza o l’Origine Etnica*, Unar, Roma

– UNICRI (ed. 2014), *Investigating and reporting online hate speech. Training manual.* , Light On Project, Turin.