Brain-be 2.0

PILLAR 3

# STATE OF THE ART

## @ntidote

Cyberviolence: defining borders on permissibility and accountability

Promotor(s)
Michel Walrave, University of Antwerp
Catherine Van de Heyning, University of Antwerp
Vanessa Fransen, University of Liège
Cécile Mathys, University of Liège
Jogchum Vrielink, University Saint-Louis

**Keywords**

Digital services
- Internet service provider
- E-commerce
- Fundamental rights
- Anti-discrimination
- Image-based abuse

University of Antwerp
LIÈGE université
UNIVERSITÉ SAINT-LOUIS BRUXELLES
belspo

## Introduction

The @ntidote project focuses on the use of social media as a platform of content that may constitute cyberviolence, in particular the posting and distribution of online hate speech ('online hate') and the non-consensual distribution of intimate images ('NCII'). ISPs are already acting against this harmful online content, but the current approach is often considered unsatisfactory. Therefore, NGOs, national and EU authorities currently discuss whether the exemption of liability for ISPs in the light of cyberviolence is still justified and whether authorities should not step in further. The proposed project examines to what extent the current approach on cyberviolence is effective or needs further improvement. It focuses on two types of cyberviolence, i.e. online hate speech and non-consensual distribution of intimate images (NCII). The project aims both at scrutinising the adequacy of the legal and judicial tools in Belgian and the European Union to better fight these forms of cyberviolence, as well as improving the understanding of these online behaviours and their prevalence in Belgium. The @ntidote project aims at providing answers as to what constitutes online harmful behaviour and the correct tools to fight these forms of cyberviolence within a fundamental rights context.

## State of the art

- ***Key research questions in the field of the research project***

The central research question of the @ntidote project is: "What action is needed and appropriate to address harmful online hate speech and non-consensual distribution of intimate images?". To answer this question, three research lines are defined to provide the required data and analysis:

- **Research question 1**: what behaviour, and based on which criteria, are considered harmful online content?

► Follow-up question: what action is considered appropriate by the population to tackle such behaviour?

- **Research question RL2**: what behaviour, and based on what criteria, is considered impermissible online content?

► Follow-up question: in so far content is considered impermissible online hate or NCII, what action is most appropriate and who should take this action (including questions on where liability for non-permissible action should be)?

- **Research question RL3**: what is the nature of the harm inflicted by online content and which coping mechanisms and actions are applied by victims?

► Follow-up question: is the current arsenal of legal / self-regulatory / technical measures considered as effective by victims to tackle the harm experienced?

- ***Main findings of past research & existing gaps of past research***

Along with the increased popularity of social media, also hate messages and NCII surged online (Waseem et al 2017). Particular to these behaviours is that they target specific individuals or groups online (Fortuna and Nunes 2018) which can result in harm for the affected users (Wulczyn, 2017). This harm consists among other of the affected individuals or groups leaving social media, silencing, or suffering personal online and offline harassment with an emotional, psychological or even physical impact (Bowler 2015, Moule 2017, Bates 2017). E.g. previous numbers from a 2014 US survey indicated that 51% of victims of NCII had experienced suicidal thoughts (Cyber Civil Rights Initiative, End revenge porn). Alarmed by this impact along with the increased popularity of social media, policy makers and ISPs have put in place new legislation, policies and guidelines on NCII and/or online hate.  Further, they introduced new procedures and technical mechanisms to prevent, detect and remove illegal online content. Yet, this has not stopped the surge of these behaviours online, begging the question whether more should be done. The EU is currently reviewing the regulatory (both binding as self-regulatory) framework on the responsibilities of information society services concerning illegal content online (Digital Service Act package). The Belgian legislator discusses changing the constitutional provision applicable to press crimes in view of online hate speech.

1) The first objective of @ntidote is to strengthen the qualitative understanding of online hate and NCII (WP1/3). This research focuses on the understanding among digital natives (selected population 15 – 25 year olds - Prensky 2001) of what constitutes online hate and NCII, what behaviour they assess as harmful / unharmful and their participation in these behaviours (perpetrator, bystander and victim). The determination of offensive content as harmful or unharmful is not clear-cut. E.g. previous online research on intimate image sharing (sexting) suggests that such behaviour often fits within individual's relational and sexual development (Van Ouytsel, 2018). The present research project therefore, intends to discern the criteria explaining the determination of online behaviour as harmful / unharmful, including age, gender, sexual orientation and culturally diverse background as potential criteria.

2) The second objective of the project is to determine what constitutes illegal online hateful speech and NCII based on the current legal framework, doctrine and case law (WP2). Definitions of online hate and NCII previously proposed by policy makers and scholars are often contested as too wide-ranging or too narrow (Gagliardone et al, 2015). The project intends to map the several (national and supranational) legal regimes in Belgium that can be applied to these online behaviours, the scope of these legal norms in addressing the several manifestations of such behaviour and the concrete application of the rules in case law do find online content legal/illegal. Previous research in other jurisdictions already suggested that tackling harmful online hate and NCII requires a varied legal framework to address the different manifestations of these behaviours (Kirchengast & Crofts, 2019; Titley et al, 2014; Ryan, 2018). However, such research is absent for the current legal Belgian framework (for NCII building further on Beyens & Lievens 2016).

3) The third objective is to develop the normative framework for removing or sanctioning online hate and NCII (WP2, 4, 5). This research develops the rationales to refrain from intervening online or to apply alternatives to a criminalised approach to cyberviolence, in particular online hate and NCII. This study includes the mapping of the normative framework that necessitates caution when intervening online, in particular the protection of freedom of speech and information. Whereas most literature is available on this framework within the US context (Kitchen 2017; Beausoleil 2019) or on the supranational level (Beliveau 2018 on the US – Council of Europe differences), little research has been conducted from the Belgian constitutional approach (Vrielink 2019). This normative framework will be applied to examine the compatibility with legal, self-regulatory and technical procedures and measures introduced or proposed by policymakers, scholars and ISPs to tackle online hate or NCII.

4) The fourth objective is to address the role of internet service providers as first responders to online hate and NCII. ISPs assess online content as permissible /impermissible on the basis of their own guidelines and policy's. In addition, an EU self-regulatory framework is in place to stimulate acting against cyberviolence, including online hate and NCII. The actions of the social media platforms are essential as to what content can be posted, will be viewed, distributed, removed or altered. Previous research concluded that harmful behaviours online can only be tackled effectively when obligations are imposed on social media to act against such content and/or a cooperative relation between authorities and ISPs are in place, in addition to acting against the perpetrator (Suzor et al, 2017). The project further examines the application of the self-regulatory and legal framework by moderators in concrete cases of online hate and NCII. As trained first responders, the research will identify how moderators assess such content as (im)permissible seeking to explain variations.

5) The fifth and final objective of @ntidote is to deepen the understanding of harm and other victim experiences. The research builds on previous research discerning several consequences of cyberviolence for victims (a.o. Bowler 2015; Moule 2017; Bates 2017). The project intends to assess whether age, gender, sexual orientation, and culturally diverse background affects the appreciation of harm caused by these behaviours as well as victims' coping mechanisms. Further, victims' actions are examined to understand why they take action (and what action) against harmful content and the hindrances they encounter.

- ***New research contributions***

First, the @ntidote project intends to supplement ongoing international research with new qualitative and quantitative data on the understanding and prevalence of online hate and NCII, adding new perspectives to the ongoing research on these behaviours (i.e. appreciation on the basis of gender, sexual orientation and culturally diverse background) and include a new approach combining a legal and social sciences research methodology.

Second, the project will add a Belgian and European approach to the normative discussion on whether and how to tackle cyberviolence to the ongoing mostly US and Australia centred research, including innovative application to technical mechanisms to reacting to / pre-empting such content.

Third, @ntidote will start a new and mainly under-researched line of research on moderators' appreciation of online content as (im)permissible and the criteria for such assessment.

Fourth, the @ntidote project will enrich current scholarly understanding of harm inflicted by online content by adding new criteria for the impact on users (gender, sexual orientation and culturally diverse background) as well as the appreciation of victims of the effectiveness of the current arsenal of remedies.

- ***Discussion on what is expected in terms of policy maker recommendations***

The project will feed policy makers and administration with further understanding of cyberviolence for expertise-based decision-making at the national level with regard to the ongoing discussion on online press crimes in the constitution, regional initiatives, the new Digital Services Act at EU-level.

Further, the data on prevalence of behaviour and criteria for harmful behaviour (including criteria such as age, gender, sexual orientation and cultural background) will allow a more targeted approach to prevention, protection and capacity building.

Moreover, the different perspectives on what content should be tolerated / prevented / tackled and what the responsibilities of different actors allow policy makers (both political as administrative) to assess what and how cyberviolence is best targeted, and the legal and procedural gaps that are present in the current framework.

Finally, clear policy recommendation will be developed in the final report as to the future legal framework but also approach to be taken to these phenomena in order to limit harm caused by the online behaviour.

- ***Bibliographic overview***

Agnew, R, (2014). General Strain Theory. In: G. Bruinsma and D. Weisburd (eds.) Encyclopedia of Criminology and Criminal Justice, Springer New York, 1892-1900; Ajzen, I. (1991). The theory of planned behaviour. Organizational behaviour and human decision processes, 50(2), 179-211; Brennan, K.A., C.L. Clark, and P.R. Shaver, Self-report measurement of adult attachment. Attachment theory and close relationships, 1998, 46-76; Collins, K. M. T., Onwuegbuzie, A. J., & Jiao, Q. G. (2007). A Mixed Methods Investigation of Mixed Methods Sampling Designs in Social and Health Science Research. Journal of Mixed Methods Research, 1(3), 267–294; COE, T-CY Mapping Study on Cyberviolence: recommendations (9 July 2018, Strasbourg) (T-CY(2017)10); Creswell, J., Plano Clark, V., Gutmann, M., & Hanson, W. (2003). Advanced mixed methods research designs. In: Handbook of mixed methods in social and behavioural research, Sage Publications, London, 209-240; Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In: Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM), 512–515, Atlanta, ICWSM; Gagliardone, Iginio. (2015). Countering Online Hate Speech – UNESCO; Gibbons, F., M. Ouellette, J. & Burzette, R. (1998) Cognitive antecedents to adolescent health risk: discriminating between behavioural intention and behavioural willingness, Psychology and Health, 13(2), 319-339; Helsper, E. & Eynon, R. (2009) Digital natives: where is the evidence?, British Educational Research Journal 36(3), 503 – 520; Kesharwani, A. (2020) Do (how) digital natives adopt a new technology differently than digital immigrants? A longitudinal study, Information & Management 57(2), 103170, 1 – 16; Kidd, S. A., & Kral, M. J. (2005). Practicing participatory action research. Journal of Counseling Psychology, 52(2), 187-195; King, G. & Wand, J. (2007) Comparing incomparable survey responses: Evaluating and selecting anchoring vignettes. Political Analysis, 15 (1), 46-66; Langer, P. (2016) The Research Vignette: Reflexive Writing as Interpretative Representation of Qualitative Inquiry--A Methodological Proposition, Qualitative Inquiry, 22(9), 735-744; Noguiera dos Santos, C, Melnyk, I & Padhi I (2018) 'Fighting Offensive Language on Social Media with Unsupervised Text Style Transfer', ACL comments, arXiv:1805.07685: Prensky, M. (2001), Digital Natives, Digital Immigrants Part 1, On the Horizon, 9(5), 1-6; Roger, D., Jarvis, G., & Najarian, B. (1993). Detachment and coping: The construction and validation of a new scale for measuring coping strategies. Personality and Individual Differences, 15(6), 619-626; Shaw L (2011) Hate Speech in Cyberspace: Bitterness without Boundaries. Notre Dame J.L. Ethics & Pub Pol'y 25(1), 279 – 304; Suzor, N., Seignior, B. & Singleton, J. (2017). Non-consensual porn and the responsibilities of online intermediaries, Melbourne University Law Review 40(3), 1057-1097; Ullmann, S. & Tomalin, M. (2020) Quarantining online hate speech: technical and ethical perspectives, Ethics and Information technology 22, 69-80; Van Ouytsel J., Punyanunt-Carter, N.M., Walrave M. & Ponnet, K. (2020) Sexting within Young Adults' Dating and Romantic Relationships. Current Opinion in Psychology, 36 (December) 55–59; 5; Walker, S., Sanci, L. & Temple-Smith, M. (2013) Sexting: Young Women's and Men's Views on Its Nature and Origins. Journal of Adolescent Health, 52(6), 697-701