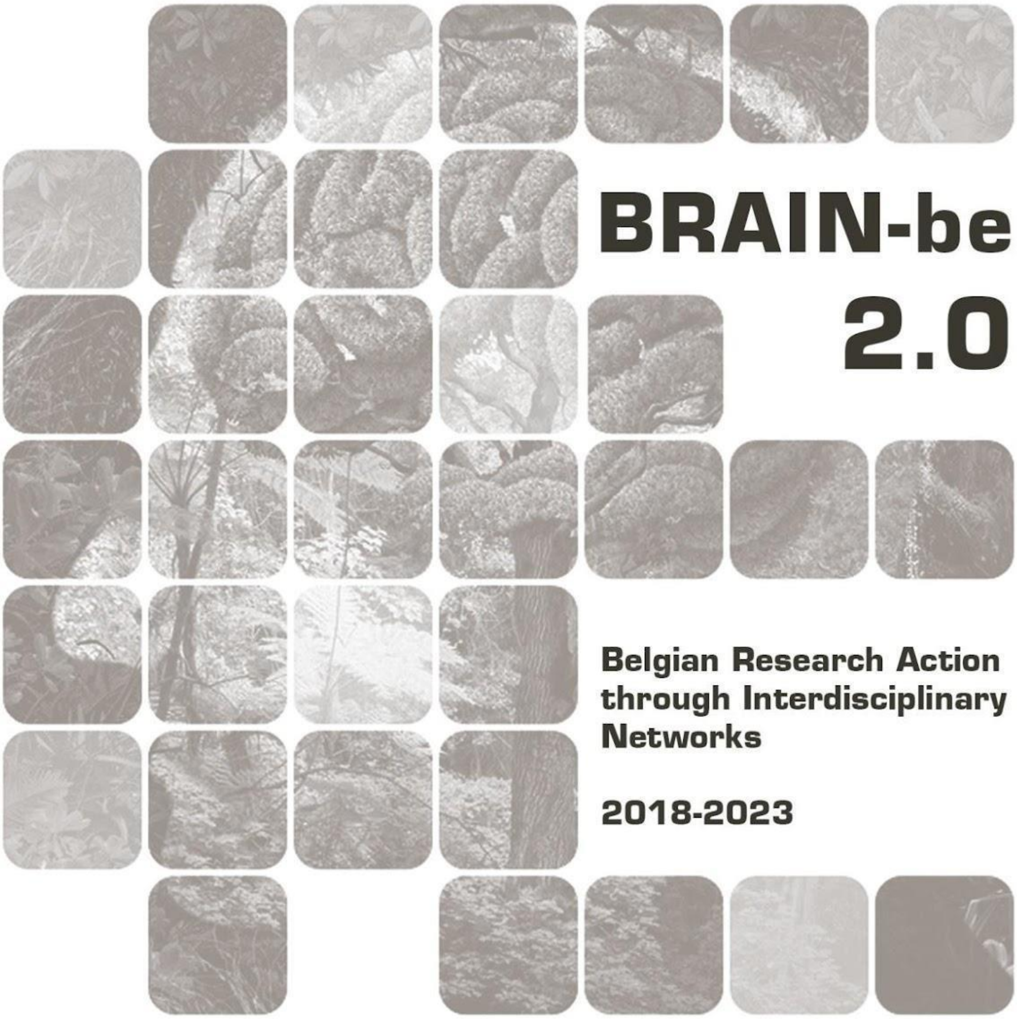


Cyberviolence: defining borders on permissibility and accountability - @ntidote 2.0

Michel Walrave, University of Antwerp
Catherine Van de Heyning, University of Antwerp
Vanessa Franssen, University of Liège
Cécile Mathys, University of Liège.
Jogchum Vrielink, University UCLouvain - Saint-Louis-Bruxelles
Mona Giacometti, University of Antwerp
Aurélie Gilen, University of Antwerp
Océane Gangi, University of Liège

BRAIN-be
2.0

Belgian Research Action
through Interdisciplinary
Networks

2018-2023

NETWORK PROJECT

Cyberviolence: defining borders on permissibility and accountability - @ntidote 2.0**Contract - B2 /202 / P3 /@ntidote 2.0.**

FINAL REPORT

PROMOTORS:

Michel Walrave, University of Antwerp (Coordinator)
Catherine Van de Heyning, University of Antwerp
Vanessa Franssen, University of Liège
Cécile Mathys, University of Liège
Jogchum Vrielink, University Saint-Louis-Bruxelles

AUTHORS:

Michel Walrave, University of Antwerp
Catherine Van de Heyning, University of Antwerp
Vanessa Franssen, University of Liège
Cécile Mathys, University of Liège
Jogchum Vrielink, University of UCLoucain - Saint-Louis-Bruxelles
Mona Giacometti, University of Antwerp
Aurélie Gilen, University of Antwerp
Océane Gangi, University of Liège

Published in 2023 by the Belgian Science Policy Office
WTCIII
Simon Bolivarlaan 30 bus 7
Boulevard Simon Bolivar 30 bte 7
B-1000 Brussels
Belgium
Tel: +32 (0)2 238 34 11
http://www.belspo.be
http://www.belspo.be/brain-be

Contact person: Aziz Naji
Tel: +32 (0)2 238 36 46

Neither the Belgian Science Policy Office nor any person acting on behalf of the Belgian Science Policy Office is responsible for the use which might be made of the following information. The authors are responsible for the content.

M. Walrave, V. Franssen, C. Mathys, C. Van de Heyning, J. Vrielink, M. Giacometti, O. Gangi A. Gilen, ***Cyberviolence: defining borders on permissibility and accountability - @ntidote 2.0,*** Final Report. Brussels: Belgian Science Policy Office 2023 – 146 p. (BRAIN-be 2.0 - (Belgian Research Action through Interdisciplinary Networks)

TABLE OF CONTENTS

ACRONYMS & ABBREVIATIONS

AI	Artificial intelligence
CEDAW	Convention on the Elimination of all Forms of Discrimination against Women
CERD	Committee on the Elimination of Racial Discrimination
CJEU	Court of Justice of the European Union
COE	Council of Europe
CSAM	Child Sexual Abuse Material (also: CSEM – Child Sexual Exploitation Material)
DSA	Digital Services Act
ECHR	European Convention of Human Rights
ECtHR	European Court of Human Rights
EU	European Union
FRA	Fundamental Rights Agency of the European Union
GBV	Gender-Based Violence
IBSA	Image-Based Sexual Abuse
ICERD	International Convention on the Elimination of All Forms of Racial Discrimination
ICT	Information and Communication Technology
IEWM	Institute for the Equality of Women and Men
LEA	Law Enforcement Authority
NCII	Non-Consensual Dissemination of Intimate Images
OHS	Online Hate Speech
OSP	Online Service Provider
MS	Member State
PTSD	Post Traumatic Stress Disorder
PWM	Prototype Willingness Model
SEM	Structural Equation Model
T&C	Terms & Conditions (also: Terms of Use or Terms of Service)
VLOP	Very Large Online Platform
U.S.	United States of America

ABSTRACT

1. Context

@ntidote is a multi-disciplinary research project that aims to develop an antidote for two types of cyberviolence: online hate speech (OHS) and non-consensual dissemination of intimate images (NCII). The team has more specifically investigated these behaviours among adolescents and emerging adults (between 15 and 25 years old). It has approached the phenomena from the perspective of social sciences, criminology, anthropology, and legal sciences.

2. Objectives

The overall objective of the @ntidote project is to better understand these manifestations of cyberviolence and evaluate whether the current approaches to tackle these online behaviours are effective. Apart from the scientific output, the project intends to equip policy makers, OSPs, and civil society with new data to fill the current gaps in our scientific knowledge on the prevalence of OHS and NCII among digital natives. It is also designed to improve the understanding of OHS and NCII in order to decide on the delineation of (il)legal or (im)permissible content as well as to shape the appropriate framework to address these behaviours.

3. Conclusions

The @ntidote project finds that adolescents and emerging adults regularly encounter OHS and NCII online. Whereas adolescents and emerging adults are often confronted with these behaviours, there is clearly a wide variety in understanding of what constitutes OHS and NCII. Age, sexual orientation (for OHS) and ethnicity are relevant criteria for higher levels of victimisation. Victims report a an emotive impact of these behaviours. The research further showed that notwithstanding the substantial impact, digital natives will not easily contact law enforcement or victim support organisations. Filing a criminal complaint might also not be the most effective step, as complaints of OHS and NCII are often discontinued due to procedural reasons, capacity, or prioritisation. Conclusions also show a wide variety in rules and procedures applied by online service providers in removing OHS and NCII as well as in collaborating with victim support organisations. The study concludes with recommendations as to media literacy, legal framework, enforcement, victim support and research as antidotes to OHS and NCII.

4. Keywords

Cyberviolence, digital natives, online service providers, online hate speech, non-consensual dissemination of intimate images, victims.

1. INTRODUCTION

"In the effort to maximize the benefit and minimize the harms of social media on children, we have not made enough progress. As a consequence, I worry about the mental health and well-being of our children".

In May 2023, Time Magazine quoted the U.S. Surgeon General Dr. Vivek Murthy on his findings regarding the impact of social media on children's well-being after having issued an Advisory on Social Media and Youth Mental Health (Park, 2023). The U.S. Surgeon General argued that minors were too often confronted with harmful content impacting their overall well-being. He added that he had witnessed the serious impact of online harmful content on minors not just in the U.S., but also in other countries he visited and studied.

The concern for the impact of harmful content on digital natives is at the core of the BELSPO @ntidote project. This project focused specifically on two behaviours of unlawful conduct, namely online hate speech (OHS) and the non-consensual dissemination of intimate images (NCII). OHS and NCII are concrete behaviours of the broader category of cyberviolence, i.e. the use of computer systems to cause, facilitate, or threaten with violence against individuals that results in, or is likely to result in, physical, sexual, psychological or economic harm or suffering. Moreover, cyberviolence may include the exploitation of the individual's circumstances, characteristics, or vulnerabilities (COE T-CY's Mapping Study on Cyberviolence, 2018). The purpose of the study is to arrive at concrete recommendations – or antidotes – against these online harmful behaviours to be used by Belgian (and European) policymakers and stakeholders, such as child and youth protection organisations, safer internet and media literacy centres, education, media, OSPs and other organisations that sensitise adolescents and emerging adults concerning digital media use.

This study starts from the findings in previous research on how social media and other digital communication applications (such as instant messaging apps) have revolutionised the way we communicate, look for information, and interact (Döring, 2009; Walrave et al., 2015). These technologies allow us to share experiences in text and pictures with a broad audience in just a split-second (Villanti et al., 2017). Moreover, because of the global reach of cyberspace, social media and communication apps enable us to share our experiences and live those of others beyond any physical boundaries. The impact is particularly felt by the Gen Z generation, as they were raised with social media as true digital natives (Margaryan et al., 2011). Social media can play an important role in young people's development and contribute to the self-presentation and -development of adolescents and emerging adults. Online, they can meet peers, discover, and experiment without virtually any limits. Further, avatars and other methods to ensure online anonymity and privacy allow young people to develop their identity and interact with others in an anonymous - and therefore often considered a safe - setting.

However, social media and associated communication apps are well-reported to have a serious downside. Harmful and often unlawful content is rampant on the Internet. Given that adolescents and emerging adults spend considerable time on social media and direct communication apps, they are often confronted with harmful content. The impulsive, fast, and unpredictable nature of the Internet also fuels various forms of violence, including cyberbullying, online harassment, cvyber dating abuse, OHS and NCII. In the framework of this research project, the team decided to focus on these last two forms of cyberviolence, as they have been under-researched in Belgium, whereas international research shows these forms of harmful online behaviours are very common among young people.

To further address these two forms of cyberviolence, it is important to conceptualise the behaviours properly. As to online hate speech, it is not easy to find a consensus of what this online behaviour exactly entails, as there is no legal definition of OHS nor is there a common understanding within the literature on the notion. A common denominator in literature to define OHS is that it includes an expression through the use of internet which targets a person or a group of persons based on a personal characteristic or status, such as gender, skin colour or religion. Moreover, the purpose of OHS is to express intense antipathy, to disrespect, even to harm, or to gain social status (Burch, 2018).

As to non-consensual dissemination of intimate images, there is a basis to find consensus by the implementation of a definition of NCII in the Belgian Criminal Code. In the Criminal Code, NCII is defined as displaying, making accessible or disseminating visual or audio content of a naked person or a person performing an explicit sexual act without their consent or without their knowledge, even if that person has consented to its creation. In literature, NCII is considered a particularly serious form of image-based sexual abuse (IBSA). For NCII to constitute a form of cyberviolence, it needs to take place on or via the internet.

Both behaviours (OHS and NCII) are regarded as forms of online violence because they can cause psychological, physical, sexual and/or social and economic harm to victims (Cookingham & Ryan, 2015; Saha et al., 2019; Singh, 2021). The impact of these behaviours is profound. Victims have reported experiencing feelings of guilt, shame, sadness, and frustration when affected by OHS or NCII (Van de Maele et al., 2023 ; Wachs et al., 2022). In the light of highly mediatised cases in Belgium on OHS and NCII affecting minors and young adults, there is an increased attention for these forms of cyberviolence. Nevertheless, there have been legislative initiatives to better regulate OHS and NCII, support organisations have developed further victim support and communication campaigns for stakeholders (e.g. young people, parents, teachers), and scholarly research into these behaviours is burgeoning. During recent years, several initiatives have been taken to prevent victimisation and perpetration via information campaigns in schools and focusing on media literacy among young people.

Yet, despite all these important efforts, it appears that insufficient progress is being made in improving the status quo of OHS and NCII. Serious cases of OHS and NCII continue to be reported by media and research. It begs the question of how to change this culture of cyberviolence. Moreover, when prevention fails, it is even more important that victims are supported. Yet, victims support organisations as Unia, the Institute for Equality of Women and Men and Child Focus have already highlighted that they struggle to respond to all complaints and incoming questions for advice due to budgetary and capacity restrictions.

As cyberviolence is enacted on online platforms or via direct messaging applications that are privately owned, mostly by foreign major companies, the reach of national law and public policy only goes so far. The willingness and capabilities of these companies (online service operators or OSPs) to prevent, find, and remove harmful content is essential in tackling cyberviolence. Several of those companies have stepped up by improving their terms of services, sanctioning users that post OHS or disseminate NCII, and by removing such content. The cooperation between OSPs and national support organisations that report OHS or NCII has proven particularly fruitful in tackling OHS and NCII. Yet, research, lived victim experiences, and support organisations' analysis show that much more needs to be done and that the actions of OSPs are often incoherent, too late, and too limited.

The above shows that there is still insufficient knowledge on the prevalence of OHS and NCII among young people in Belgium, the impact of these behaviours on victims, the follow-up of incidences of OHS and NCII by law enforcement, justice and support organisations, and the role of OSPs in tackling cyberviolence in Belgium. The @ntidote-project therefore focused on 5 aspects of OHS and NCII in Belgium, namely the qualitative understanding among young people (i), the legal framework delineating lawful from unlawful content (ii), the prevalence of OHS and NCII among young people in Belgium, including their understanding of harmful and unharmful content (iii), the position of OSPs and their appreciation of content as permissible or non-permissible (iv), and finally the coping mechanisms and support needs of victims both from the perspective of victims themselves as of support organisations (v).

The research is intrinsically interdisciplinary: the @ntidote team consisted of researchers specialised in sociology, sexology, criminology, law and anthropology. By bridging the differing expertises via interconnected work methods, the research resulted in novel and promising results and recommendations. Preventing and curing the impact of OHS and NCII is a massive and global challenge that will not be tackled overnight and by one nation alone. However, the @ntidote team is convinced that the recommendations that follow from the research are worth considering and investing in to protect the mental health and well-being of adolescents and emerging adults.

2. STATE OF THE ART AND OBJECTIVES

Along with the increased popularity of social media, 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 things, 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). For example: 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 OSPs have put in place new legislation, policies and guidelines on NCII and/or OHS. Furthermore, they introduced new procedures and technical mechanisms to prevent, detect and remove illegal online content. However, this has not stopped the surge of these behaviours online, posing the question whether more should be done. The EU has just adopted a new regulatory framework on the responsibilities of OSPs concerning illegal content online (The Digital Service Act) and is discussing new substantive rules on cyberviolence (particularly gender-based cyberviolence, or GBV, and hate crimes). The Belgian legislator also repeatedly discussed changing the constitutional provision applicable to press crimes in view of prosecuting online hate speech more effectively.

The overall objective of the @ntidote research project is to equip policy makers, public authorities, internet service providers, and civil society with the required data on and understanding of OHS and NCII to decide on the delineation of legal and illegal content as well as the appropriate framework to address these behaviours. Firstly, it produces new data for Belgium to fill the current gaps in our scientific knowledge on the prevalence of OHS and NCII among digital natives. Moreover, the project enhances the understanding of the relevance of criteria such as age, gender, sexual orientation, and culturally diverse backgrounds and of victims' coping strategies. Secondly, it maps the legal and technical instruments to fight illegal content and identify the hindrances to action for victims. In doing so, the project provides a framework for further research on prevention and the required future capacities needed to tackle harm inflicted online. Moreover, this research strengthens civil society with scientific insight to debate the limits of online speech and content. Finally, the research also contributes to the current debate on rethinking the role and liabilities of the digital economy, in particular OSPs, for online harmful content. Insight into their users' appreciation of content as harmful and existing procedures, can indeed contribute to further development of social media.

The overall ambition of the project has been split into five more specific objectives that the team has developed through the design of several work packages (WP), which were built upon five objectives.

The first objective is to strengthen the qualitative understanding of OHS and NCII (WP1/3). This research focuses on the understanding among digital natives (selected population between 15 and 25 years old) of what constitutes online hate and NCII, what behaviour they assess as harmful and unharmful and their participation in these behaviours (perpetrator, bystander and victim). The term "digital natives" refers to the first generation that has grown up with the expansion of the Internet (Bennett et al., 2008; Prensky, 2001). According to Prensky's study, the digital experiences of digital natives differ in several ways from those of previous generations, referred to as digital immigrants (2001). Those young people are widely recognised for their extensive use of social media and various online platforms (Costello et al., 2016). This digital engagement offers them opportunities to develop diverse digital skills, including technical proficiency in using social networks and algorithms, creative expression through sharing photos and videos, and improved social and communicative abilities (Aranda Juárez et al., 2020).

However, the digital natives are also widely recognised as being particularly vulnerable to cyberviolence (Costello & Hawdon, 2020). Moreover, certain individual characteristics have been identified as being associated with online victimisation, such as sexual orientation (Baider, 2019), gender (Döring & Mohseni, 2019), and cultural background (Ortiz, 2021).

The determination of offensive content as harmful or unharmful by the focus group of digital natives is not clear-cut. For example, previous online research on consensual intimate image sharing (sexting) suggests that such behaviour often fits within individuals' relational and sexual development (Van Ouytsel, 2018). Other studies have indicated that NCII may be perceived as less harmful for men than women (Dekker et al., 2019), and the level of harm may depend on whether the initial image sharing was consensual or not (Dekker et al., 2019). Concerning OHS, the consequences depend on how victims perceive it, influenced by factors such as the perpetrator's identity, content, and the targeted individuals (Chetty & Alathur, 2018). Some researchers argue that hate speech might not necessarily cause harm, as intention of perpetrator and the content diffused are prioritised (Chetty & Alathur, 2018). The present research project intends to further develop the harmfulness of those behaviours by discerning the criteria explaining the determination of online behaviour as harmful or unharmful, including gender, sexual orientation, and culturally diverse background as potential criteria, and by including the perspectives not only from the victim but also from the perpetrator and bystanders.

The second objective of the project is to determine what constitutes illegal OHS 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 the law does find certain online content legal or illegal. Previous research in other jurisdictions suggests 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).

The third objective is to develop the normative framework for removing or sanctioning OHS and NCII (WP2, 4, 5). 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 OSPs to tackle online hate or NCII.

The fourth objective is to address the role of OSPs as first responders to online hate and NCII (WP 4). OSPs assess online content as permissible or impermissible based on their own guidelines and policies. In addition, an EU self-regulatory framework is in place, and has been recently further developed with the adoption of DSA to stimulate acting against cyberviolence, including OHS 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 OSPs is 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 identifies how moderators assess such content as (im)permissible, seeking to explain variations.

The fifth and final objective of @ntidote is to investigate the negative emotions of victims of NCII and OHS, and as such the harm inflicted by these behaviours (Bowler, 2015; Moule, 2017; Bates 2017). Moreover, the role of victim support organisations is evaluated from both digital natives' and the organisations' perspective. Organisations were interviewed to further substantiate the research with information about victim experiences, prevalence rates, their method to tackle these behaviours, and suggestions for future improvement.

3. METHODOLOGY & RESULTS

3.1 QUALITATIVE UNDERSTANDING OF OHS AND NCII

3.1.1 METHODOLOGY

a. Research questions

The main objective of WP1 is to enhance the qualitative comprehension of online hate speech and NCII of a population of interest. This research primarily focuses on digital natives, a population with distinct characteristics in terms of their utilisation of the digital world and their relationship with it. The term "digital natives" refers to the first generation that has grown up with the expansion of the Internet (Bennett et al., 2008; Prensky, 2001). According to Prensky's study, the digital experiences of digital natives differ in several ways from those of previous generations, referred to as digital immigrants (2001). This includes the role of digital platforms in peer communication (Keipi et al., 2017). Given the technological advancements around the 2000s (Keipi et al., 2017), this research focuses on individuals born between 1997 and 2007. This aspect of the study examines three elements: (i) their definition of online hate speech and NCII, (ii) what these young individuals perceive as harmful or unharmful, and (iii) the multiple status involved as perpetrators, bystanders, and victims and the perceived motives to harm. The presence of diversity variables or individual characteristics (such as gender, sexual orientation and cultural background) will be integrated into this study to understand these elements among various profiles. Please notice that the results for the individual characteristics focus only on online hate speech.

i. Digital natives: why is this a population of interest?

As mentioned before, young individuals aged 15 to 25 are widely recognised for their extensive use of social media and various online platforms (Costello et al., 2016). This digital engagement offers them opportunities to develop diverse digital skills, including technical proficiency in using social networks and algorithms, creative expression through sharing photos and videos, and improved social and communication abilities (Aranda Juárez et al., 2020). However, the extent to which these skills are fully realised remains a topic of discussion in literature (Estanyol et al., 2023).

Moreover, the challenges faced by adolescents and emerging adults are reflected in their online behaviour. Their virtual interactions align with the social developmental needs of this age group, as they explore various online communities to forge their identities and assert their individuality (Cannard, 2019). Through this exploration, they experiment with boundaries of privacy and self-disclosure, all without direct physical contact (Tisseron, 2011).

Considering the digital environment's unique characteristics, these same mechanisms are facilitated by technical features, which fosters the formation of identity reinforcement bubbles and a sense of group belonging (Keipi et al., 2017). Additionally, the Social Identity Model of Deindividuation Effects (SIDE) suggests that the digital context and anonymity may place a stronger emphasis on elements of social identity rather than personal identity. Consequently, individuals may be more inclined to present themselves and adjust their online behaviour based on their group affiliations, rather than emphasizing their individual identity. This dynamic further intensifies the comparison between different groups within the digital space. Related to this, the high presence on social media and the search for identity makes the digital natives particularly exposed to online hate speech (Bautista-Ortuño et al., 2018; Costello & Hawdon, 2020; Hawdon et al., 2017; Keipi et al., 2017; al Serhan & Elareshi, 2019).

Furthermore, the phenomenon of sexting - defined as the consensual sending, receiving, and forwarding of nude, semi-nude, or sexually explicit images within digital forms of communication - among adolescents and emerging adults contributes to the concept of 'extimacy,' signifying the growing trend of publicly sharing intimate aspects of one's life (Tisseron, 2007). This practice highlights the evolving dynamics of privacy in the digital era, as young people navigate the delicate balance between safeguarding their privacy and expressing themselves freely in their online interactions.

However, the digital natives are also widely recognised as being particularly vulnerable to cyberviolence (Costello & Hawdon, 2020). Moreover, certain individual characteristics have been identified as being associated with online victimisation, such as sexual orientation (Baider, 2019), gender (Döring & Mohseni, 2019), or cultural background (Ortiz, 2021).

Belgian and international research focusing on adolescents aged 12 to 17 finds that the prevalence rates of NCII victimisation and perpetration vary between 5% and 8% and 10% and 12% respectively (Glowacz & Goblet, 2020; Madigan et al., 2018; Van Ouytsel et al., 2021). A Spanish study further indicates that the risks can appear as early as the age of 12 (Gámez-Guadix et al., 2022). In addition, more than a quarter of university students reported non-consensually sharing a sexually explicit message or image with close friends when they were under 18 years old (Patel & Roesch, 2020). Young adults aged 20 to 29 also become victims more frequently, with LGBTQ individuals being at higher risk, as well as individuals from diverse cultural backgrounds (Henry et al., 2019). Studies also point out that men are more likely to agree with sharing sexually explicit images of their girlfriends (Walker & Sleath, 2017).

Adolescents and emerging adults appear to be more susceptible to being exposed to online hate speech due to their online presence (Castaño-Pulgarín et al., 2020; Costello et al., 2019; Hawdon et al., 2017). A recent study focused on the exposure to online hate speech among individuals aged 18 to 25 in six distinct countries (Finland, France, Poland, Spain, the United Kingdom, and the United States) (Reichelmann et al., 2021). Results showed that most participants had seen online hate speech within the preceding three months. France and the United Kingdom had significantly lower rates than the other nations. Notably, several differences exist between countries concerning the samples and emotional reactions. The most prominent emotion among respondents was as follows: sadness in Finland, Poland and the United States, and anger in France, Spain, and the United Kingdom. The emotions of anger, hatefulness, sadness, and shame were experienced by over 40% of respondents in all countries. This audience appears highly exposed to this type of content. However, it is interesting to wonder whether this is also the case for victims. While certain studies indicate an absence of a statistically significant relationship between the duration of online activity and victimisation (Costello et al., 2021), others specify that the association might indeed exist for victimisation (Siegel, 2020) or exposure (Hawdon et al., 2017). In addition, individuals who might be more prone to becoming targets of online hate speech include LGBTQIA+ people (Meyer, 2010), women (Cottee, 2021; Jane, 2018), and individuals with a foreign origin (Küpper et al., 2010). This introduces the concept of intersectionality, meaning that certain individuals possess multiple targetable characteristics (Mcphail, 2002; Ging, 2019).

ii. What are the definitions of online hate speech and NCII?

As general state of the art pointed out, online hate speech and NCII fall under the category of cyberviolence: *"encompassing behaviours that utilise computer systems to provoke, facilitate, or threaten violence against individuals"* (Crespi & Hellsten, 2022, p. 392). Cyberviolence can take various forms, including audio, video, or textual content, and are facilitated through instant messaging and social media platforms. They may involve insults, rumors, the dissemination of images, identity impersonation, or the fraudulent use of personal information (Willard, 2004). Online hate speech and NCII are two different behaviours, even if they share some similarities.

There is a lack of clarity about which behaviours fall under the concepts of "online hate speech" (e.g., do insults based on appearance qualify?) and NCII (e.g., do threats to distribute intimate images fall within this category?). Notably, studies on online hate speech lack consensus in the criminological, sociological, and psychological fields, as well as in legal texts. The definition of online hate speech varies depending on the author. For some, it involves the use of aggressive or offensive language targeting a specific group of individuals who share common characteristics, such as gender, ethnicity, beliefs, religion, or political preferences (Zhang and Luo, 2018). Others view it more broadly as expressing hatred towards a collective with the objective or consequence of exclusion (Hawdon et al., 2017). Other terms utilised in this context include hate speech (Chetty and Alathur, 2018), online hatred (Salminen et al., 2020), hateful content (Costello et al., 2019), and hate crimes (Jacks and Adler, 2016). From a legal perspective, the Council of Europe defines hate speech as "any kind of communication that promotes, incites, spreads or justifies violence, hatred, or discrimination against persons or groups, or that insults or denigrates them, on the basis of their personal characteristics or real or attributed status, including race, color, language, religion, nationality or ethnic origin, age, disability, sex, gender, gender identity, and sexual orientation." (Recommendation CM/Rec(2022)16[1] of the Committee of Ministers to Member States on combating hate speech, adopted by the Committee of Ministers on 20 May 2022, at the 132nd Session of the Committee of Ministers). The @ntidote project uses this definition in full, as online hate speech is any form of expression (texts, videos, audios, photos, images, games, and others) through the use of the Internet (digital platforms, social networks, and others), which is motivated by prejudice, intolerance, or discrimination, and targets a group of people (or an individual from this group) sharing a common inherent or acquired property or characteristic, whether actual or perceived, such as ethnic origin, belief, disability, gender, or sexual orientation. The purpose of this expression would be to convey intense antipathy, disrespect, or even harm, and/or to gain social status (Burch, 2018).

To define NCII, it is interesting to consider two related concepts: sexting and image-based sexual abuse. Sexting is widely accepted as the consensual sharing of intimate images or sexually explicit texts (Holmes et al., 2021; Wachs et al., 2021; Sparks, 2022). It is a practice used as an attempt at seduction within an intimate relationship or as a marker of trust in a friendly relationship (Cooper et al., 2016; Glowacz & Goblet, 2020). Image-based sexual abuse (IBSA) refers to the act of taking or sharing nude and sexual photographs of another person without their consent. It encompasses three primary behaviours: non-consensual taking of nude or sexual images, non-consensual sharing of nude or sexual images (referred to as NCII), and threats to share nude or sexual images. Thus, NCII, as studied in our research, is a form of IBSA. It is essential to differentiate NCII from revenge porn, as the latter is a specific form of NCII driven by vengeance between current or former partners. The literature lacks consensus on the use of the term NCII, as alternative terms such as non-consensual pornography (Sparks, 2022) or aggravated sexting (Gassó et al., 2019) can also be found. Furthermore, it is useful to specify that NCII is a term used by researchers and sometimes by practitioners, but not originated from young individuals.

iii. How do young people perceive the harmfulness of online hate speech and NCII?

Investigating the harmfulness of online hate speech and NCII is crucial for understanding and addressing these behaviours effectively. It also helps determine the appropriate psychosocial and legal actions to be taken. However, there is a scarcity of studies regarding NCII, and these few studies focus more on sexting and its consequences for victims. Few studies have explored the perspectives of bystanders and perpetrators involved in NCII. NCII has been associated with various psychosocial effects (Alonso & Romero, 2019; Sparks, 2022) and links to other forms of victimisation (Couturiaux et al., 2021; Setty, 2020). Studies have indicated that NCII may be perceived as less harmful for men than women (Dekker et al., 2019), and the level of harm may depend on whether the initial image sharing was consensual or not (Dekker et al., 2019).

With online hate speech, the consequences depend on victims' perception and are influenced by factors such as the perpetrator's identity, content, and the targeted individuals (Chetty & Alathur, 2018). Some researchers argue that hate speech might not necessarily cause harm, as the intention of the perpetrator and the content diffused are prioritised (Chetty & Alathur, 2018). Other factors influencing perceived offensiveness include the gender of the speaker, with content written by men often seen as more offensive (Bautista-Ortuño et al., 2018). Endogenous factors, such as the sense of identification with the targeted group, also play a role in determining offensiveness (Bautista-Ortuño et al., 2018; Costello et al., 2019). Indeed, individuals with strong group affiliations may perceive hate speech more intensely (Bautista-Ortuño et al., 2018). In addition, victims of online hate speech are more likely to recognise speech as hate speech (Costello et al., 2019). Finally, affiliation with groups advocating deviant norms may lead to a perception of hate speech as less offensive (Costello et al., 2019). It has furthermore been identified that the perception of offensiveness is dependent on the individual characteristics. According to literature, an intersectional approach would involve considering various characteristics based on social position, gender, sexual orientation, cultural background, etc. Intersectionality suggests that different systems of oppression, such as racism and sexism, are interconnected and cannot be hierarchically ranked (Meyer, 2010).

In relation to coping mechanisms, victims of online hate speech employ various strategies that may depend on their perception of victimisation and multiple individual factors (al Serhan & Elareshi, 2019; Wachs et al., 2020). The selected studies address online hate speech as one of the facets of cyberhate. Research suggests that girls tend to use coping mechanisms more frequently than boys (Wachs et al., 2020; al Serhan & Elareshi, 2019). The strategies employed are also influenced by the individual's age and acquired technical skills (Wachs et al., 2020). Being a victim of such content may be correlated with perpetration, especially as these contents can become normalised (Costello et al., 2020). Only 2% of cyberhate victims report it to the police, and 4% to a professional (Wachs et al., 2020). The most used coping mechanisms are preserving evidence (screenshots, conversations, etc.), paying more attention to who has access to personal data, conveying to the person that the behaviour is unacceptable, telling the person to stop, and blocking the individual (Wachs et al., 2020).

iv. Online hate speech and NCII: a perspective through multiple roles?

When a transgressive behaviour occurs, whether online or offline, different roles can be identified. Firstly, there is the perpetrator, who is the person committing the act. Secondly, there is the victim, who is the person against whom the act is directed. Finally, other individuals present can be considered as bystanders.

The first thing to be noted is that perpetration of OHS can take various forms, whether by targeting individuals with specific comments or by sharing certain content (Awan, 2014). Literature indicates several motivations for engaging in this type of speech, such as grievance, seeking power, or the desire for group inclusion (Awan, 2014; Jacks & Adler, 2016). Similarly, the motivations behind NCII are also diverse (Harper et al., 2021), including revenge, demonstrating sexual prowess, seeking entertainment, asserting control, or financial gain (Harper et al., 2021; Henry et al., 2019). Both types of behaviours show a strong presence of social motivations. Social motives are 'the psychological processes that drive people's thinking, feeling and behaviour in interactions wi th other people' (Reinders Folmer, 2016).

The second thing to be noted is that, as to victimization, several profiles can be discerned. For both OHS as NCII, the victims often possess characteristics leading to them being perceived as 'minority groups' in terms of sexual orientation, gender, and cultural background (Baider, 2019; Henry et al., 2019; Ortiz, 2021; Walker & Sleath, 2017). However, it is important to note that NCII has a much higher victimisation rate among women (Henry et al., 2019; Walker & Sleath, 2017).

Finally, it is essential to focus on the bystanders, as cyberviolence has the unique ability to increase the number of bystanders who may experience a form of victimisation through exposure to certain content. Regarding OHS, the literature is divided. Some research indicates that high exposure to hate speech online may lead to desensitization for both bystanders and victims, resulting in a diminished perception of offensiveness (Bernatzky et al., 2022; Soral et al., 2018). Thus, bystanders of such content may be more inclined to produce similar content (Bernatzky et al., 2022), as is the case with hate speech (Soral et al., 2018). Other studies suggest that this view does not hold true in the digital context and that high exposure may make individuals more inclined to define certain remarks as online hate speech (Costello et al., 2019). Moreover, considering the specificities of the digital realm and the large number of bystanders, researchers question the criteria that determine whether someone can be considered a victim (Costello & Hawdon, 2020). When it comes to NCII, fewer studies have focused on the role of bystanders. Literature highlights the role of rape myths in shaping perceptions of victim accountability and supporting the perpetrator (Dekker et al., 2019). In both online hate speech and NCII, the role of bystanders is emphasised (Henry et al., 2020; Salminen et al., 2020).

b. Methodology for the interviews

i. Sampling

Recruitment

The team implemented various methods to ensure a diversified sample and reach participants with different diversity variables. These techniques were applied to both the French and Dutch subsamples. By employing these methods, the team aimed to ensure a diverse and representative sample for our research.

First, the team designed posters tailored to the target population (annexes 2 and 3), distinguishing between two age groups: 15 to 18 years old and 18 to 25 years old. These posters were displayed in offline locations such as universities, schools, and community organisations, as well as on online platforms like Facebook and Instagram. The posters clearly outlined the research themes, target audience, contact information for the researchers (WhatsApp, SMS, email, and QR code leading to a Google Form), and practical details including interview location, duration, incentive, and ethical precautions.

The language used in the posters allowed us to capture any relevant experiences related to the project, ensuring access to a diverse population, and attracting participants with different roles (perpetrators, victims, and/or bystanders).

Second, the team established connections with reference organisations, acting as "gatekeepers", which provided access to specific and sensitive populations such as LGBTQIA+ individuals or perpetrators. A document was shared with these gatekeepers, describing the study's purpose, the protocol for connecting researchers with potential participants, researchers' commitment to participants, and the ethical precautions taken (annex 1).

Following each interview, which lasted approximately 1 hour, the team completed a participant coding document. This document included variables related to sexual orientation, gender, cultural background, age, and status (perpetrator, victim, and/or bystander). It allowed us to track the number of participants meeting the criteria for each variable and status. If certain criteria were underrepresented (e.g., insufficient number of victims), the team adjusted its recruitment strategy to meet the requirements of our sampling matrix.

Inclusion criteria

The team applied three inclusion criteria to select participants:

- Age: Only individuals between the ages of 15 and 25 were included. This age range was chosen to focus on a specific demographic group that has grown up in the digital age and is likely to have experienced online hate speech and non-consensual dissemination of intimate images (NCII).
- Diversity: To ensure a diverse sample, the team explicitly sought participants with various sexual orientations and ethnic backgrounds. By collaborating with specific organisations such as çavaria and Roze Huis, we aimed to reach individuals representing different identities and backgrounds.
- Self-reported status: the team included individuals who identified themselves as bystanders, perpetrators, or victims of online hate speech and NCII. This allowed the team to gather insights from different perspectives within the digital natives' population. We were able to obtain different types of statuses by diversifying the sampling locations: schools, Public Youth Protection Institution, residential center for young people, etc. We indicated that we were seeking all experiences related to the phenomena.

By applying these inclusion criteria, the team aimed to capture a range of experiences and perspectives related to online hate speech and NCII among young individuals.

Sampling Matrix

The sampling approach for this study is non-probabilistic, as our objective is not to create a sample that is representative of the entire population. Instead, our sampling strategy is objective and stratified, guided by the insights gained from the literature review and the expertise of the researchers. The team has identified several relevant subgroups based on factors such as gender, sexual orientation, ethnicity, and age. To guide our sampling process, the team has developed a sampling matrix that takes these factors into account. Given the exploratory nature of this qualitative study, the team adheres to the principles of saturation and diversification. Saturation refers to the point at which collecting additional data no longer reveals new insights or perspectives on the research topic. Diversification involves ensuring representation from different subgroups within the sample (Yin, 2017). To achieve diversification, the team aimed to interview individuals from the various subgroups identified in the sampling matrix. This approach allows us to capture a broad range of perspectives and experiences.

Throughout the data collection process, the team monitored the evolution of participant profiles to ensure diversity and achieve the desired variety of perspectives (following the principle of internal diversification).

By employing these principles and strategies, the team aimed to obtain a comprehensive understanding of the research topic through insights from diverse participants across various subgroups. In total, the sample is composed of 24 participants:

Table I. Description of the participants and their individual characteristics

Description		
Gender	Male	**13**
	Female	**11**
	Non-binary	**0**
	Transgender	**0**
Age	15-17	**10**
	18-25	**14**
Cultural background (by place of birth of parents)	Belgian	**14**
	European (not Belgian)	**5**
	African	**8**
	Asian	**1**
	Others (Oceania, America,..)	**0**
Sexual orientation	Heterosexual	**13**
	Homosexual	**3**
	Bisexual	**4**
	Pansexual	**1**
	Asexual	**0**
	Others	**3**
Self-reported status	NCII - Victim	**8**
	NCII - Perpetrator	**4**

	NCII - Bystander	**20**
	OHS - Victim	**13**
	OHS - Perpetrator	**3**
	OHS - Bystander	**22**
Language	Dutch-Speaking	**10**
	French-Speaking	**14**

ii. Interviews and analysis

Interview guide

The interviews were conducted using an interview guide (annexes 4, 5 and 6), which was developed based on the literature on online hate speech, NCII, and online violence (al Serhan & Elareshi, 2019; Bautista-Ortuño et al., 2018; Castano-Pulgarín et al., 2021; Costello & Hawdon, 2020; Keipi et al., 2017; Meyer, 2010; Ortiz, 2021). The main questions focused on the definitions given to online hate speech and NCII, as well as the experiences of bystanders, victims, and/or perpetrators (self-report). For example, "What is online hate speech? If it's easier for you, could you provide an example? ”; “When does nudes sharing become harmful?” and “After our discussion about your online experiences, in which role(s) do you identify the most? A victim? A perpetrator? A bystander? A combination of several roles? ” The interview guide underwent a pre-test with French and Dutch youths who shared similar characteristics to the targeted sample.

Interview process

The interviews were conducted by two junior researchers, with three interviews conducted by a final year criminology student trained by the junior researchers. They took place at the premises of the University of Antwerp or Liège, or within their residence, depending on what the respondent preferred. A standardised procedure was followed for all interviews. The participants were warmly welcomed, and the purpose of the research was explained to them. Ethical precautions, such as the option to interrupt the interview, audio recording, and consent documents, were reviewed with the participants. The duration of the interviews ranged from 30 to 90 minutes.

Ethical precautions

To ensure the ethical integrity of the research, a comprehensive research proposal was submitted to and approved by the Ethics Committee of Social Sciences and Humanities at the University of Antwerp. The committee's regulations include specific provisions regarding the inclusion of minors (aged 14 to 18). Prior to conducting interviews with these minors, their parents received an information sheet that outlined various aspects of the research, including its process, potential consequences, participants' rights, and information about support organisations. Given the sensitive nature of the interview topics, which could potentially trigger negative emotions in participants, the researchers made a point of emphasising the availability of support organisations immediately after the interviews.

Participants were informed that they could find information about these organisations on the project's website or in the provided information sheet. These measures were put in place to prioritise the well-being and ethical treatment of the participants throughout the research process.

Analysis of the data

After transcribing the interviews, an inductive analysis was applied to analyse the data. This method focuses on deriving new insights and exploring dimensions that have not been extensively researched before. Rather than fitting the interview responses into pre-existing categories, this method allows the data itself to guide the analysis process, emphasizing the richness and uniqueness of the participants' input (Blais and Martineau, 2007; Thomas, 2006).

The analysis process involved several sub steps. Each participant's responses were summarised, and multiple rounds of analysis were conducted to identify overarching topics. These topics were then further refined and grouped into final categories. This analytical approach was applied to the main themes explored during the interviews, including the perception of human rights, social networks, cyberviolence, online hate speech, and non-consensual dissemination of intimate images (NCII).

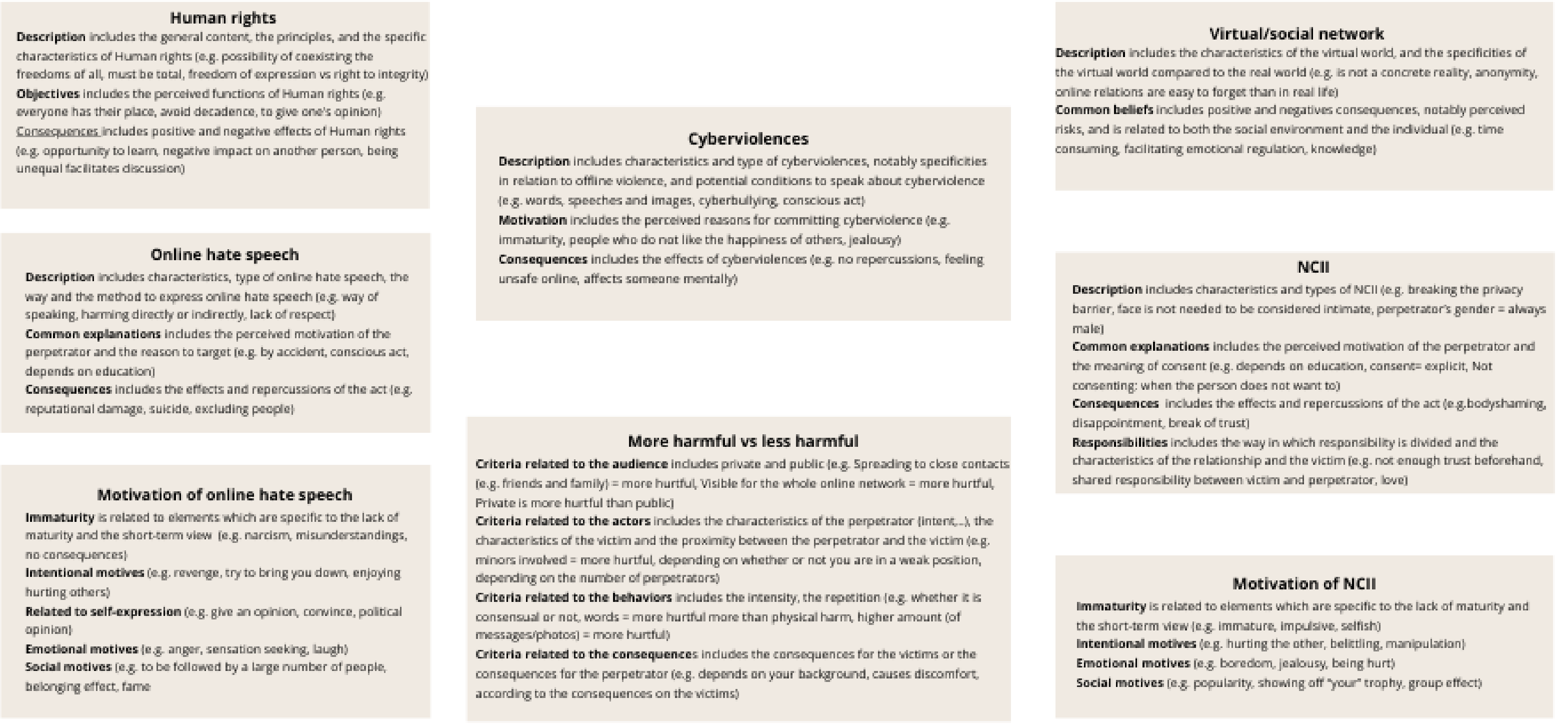
By using this method, previously unexplored categories (e.g., common beliefs) were identified, which contributed to a more comprehensive understanding of online hate speech and NCII and their contextual factors among diverse digital natives. To ensure the fidelity of the data analysis, a meeting was held on September 1st, 2022, involving the WP supervisor, the postdoctoral researcher, and the two junior researchers. During this meeting, the final categories were determined, and multiple coding techniques were applied to ensure the accuracy and consistency of the data analysis (see Figure 1). Additional analyses were also conducted to extract maximum insights from the collected data. Emerging trends and patterns in terms of definitions and experiences were identified. These categories were then integrated into a model to reveal the connections and relationships among them (see Figure 2).

Limitations

For the analysis, one needs to be mindful of the limitations of the interview guide. This guide contained numerous questions, including general cyberviolence topics, online hate speech and NCII. Due to the explanatory research, the abundance of topics covered may have contributed to a lack of clarity in the questionnaire on certain topics. This finding is a starting point for further research that could go more in-depth into certain aspects of the current research.

3.1.2 RESULTS

The inductive analysis led us to highlight eight general categories as presented in Figure 1.



Human rights

Description includes the general content, the principles, and the specific characteristics of Human rights (e.g. possibility of coexisting the freedoms of all, must be total, freedom of expression vs right to integrity)
Objectives includes the perceived functions of Human rights (e.g. everyone has their place, avoid decadence, to give one's opinion)
Consequences includes positive and negative effects of Human rights (e.g. opportunity to learn, negative impact on another person, being unequal facilitates discussion)

Online hate speech

Description includes characteristics, type of online hate speech, the way and the method to express online hate speech (e.g. way of speaking, harming directly or indirectly, lack of respect)
Common explanations includes the perceived motivation of the perpetrator and the reason to target (e.g. by accident, conscious act, depends on education)
Consequences includes the effects and repercussions of the act (e.g. reputational damage, suicide, excluding people)

Motivation of online hate speech

Immaturity is related to elements which are specific to the lack of maturity and the short-term view (e.g. narcism, misunderstandings, no consequences)
Intentional motives (e.g. revenge, try to bring you down, enjoying hurting others)
Related to self-expression (e.g. give an opinion, convince, political opinion)
Emotional motives (e.g. anger, sensation seeking, laugh)
Social motives (e.g. to be followed by a large number of people, belonging effect, fame

Cyberviolences

Description includes characteristics and type of cyberviolences, notably specificities in relation to offline violence, and potential conditions to speak about cyberviolence (e.g. words, speeches and images, cyberbullying, conscious act)
Motivation includes the perceived reasons for committing cyberviolence (e.g. immaturity, people who do not like the happiness of others, jealousy)
Consequences includes the effects of cyberviolences (e.g. no repercussions, feeling unsafe online, affects someone mentally)

More harmful vs less harmful

Criteria related to the audience includes private and public (e.g. Spreading to close contacts (e.g. friends and family) = more hurtful, Visible for the whole online network = more hurtful, Private is more hurtful than public)
Criteria related to the actors includes the characteristics of the perpetrator (intent,...), the characteristics of the victim and the proximity between the perpetrator and the victim (e.g. minors involved = more hurtful, depending on whether or not you are in a weak position, depending on the number of perpetrators)
Criteria related to the behaviors includes the intensity, the repetition (e.g. whether it is consensual or not, words = more hurtful more than physical harm, higher amount (of messages/photos) = more hurtful)
Criteria related to the consequences includes the consequences for the victims or the consequences for the perpetrator (e.g. depends on your background, causes discomfort, according to the consequences on the victims)

Virtual/social network

Description includes the characteristics of the virtual world, and the specificities of the virtual world compared to the real world (e.g. is not a concrete reality, anonymity, online relations are easy to forget than in real life)
Common beliefs includes positive and negatives consequences, notably perceived risks, and is related to both the social environment and the individual (e.g. time consuming, facilitating emotional regulation, knowledge)

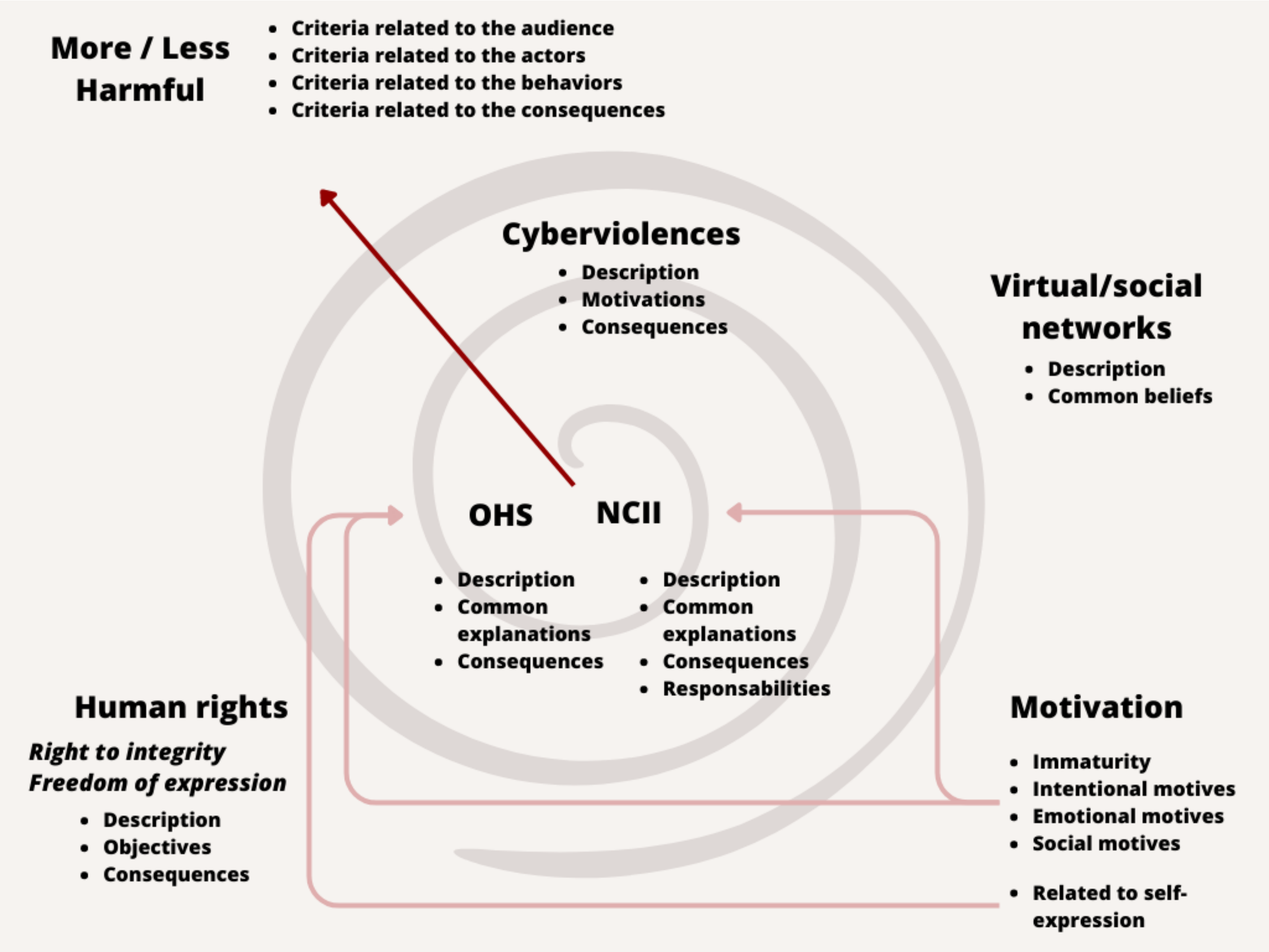
NCII

Description includes characteristics and types of NCII (e.g. breaking the privacy barrier, face is not needed to be considered intimate, perpetrator's gender = always male)
Common explanations includes the perceived motivation of the perpetrator and the meaning of consent (e.g. depends on education, consent= explicit, Not consenting: when the person does not want to)
Consequences includes the effects and repercussions of the act (e.g.bodyshaming, disappointment, break of trust)
Responsibilities includes the way in which responsibility is divided and the characteristics of the relationship and the victim (e.g. not enough trust beforehand, shared responsibility between victim and perpetrator, love)

Motivation of NCII

Immaturity is related to elements which are specific to the lack of maturity and the short-term view (e.g. immature, impulsive, selfish)
Intentional motives (e.g. hurting the other, belittling, manipulation)
Emotional motives (e.g. boredom, jealousy, being hurt)
Social motives (e.g. popularity, showing off "your" trophy, group effect)

Figure 1. Identified categories relating to the definition of OHS and NCII among digital natives



More / Less Harmful

- Criteria related to the audience
- Criteria related to the actors
- Criteria related to the behaviors
- Criteria related to the consequences

Cyberviolences

- Description
- Motivations
- Consequences

Virtual/social networks

- Description
- Common beliefs

OHS

- Description
- Common explanations
- Consequences

NCII

- Description
- Common explanations
- Consequences
- Responsabilities

Human rights

Right to integrity

Freedom of expression

- Description
- Objectives
- Consequences

Motivation

- Immaturity
- Intentional motives
- Emotional motives
- Social motives
- Related to self-expression

Figure 2. Modelling of the categories on harmfulness

a. Prevalence of OHS and NCII

The results of the analysis show a high level of prevalence for cyberviolence. The data indicate that online hate speech is a highly prevalent behaviour among adolescents and emerging adults, as 23 of the 24 respondents reported having experienced it in some way, whether as bystanders, victims, or perpetrators. The same applies to NCII, where 22 of the 24 respondents have witnessed this phenomenon. The experience appears to be prevalent among many individuals aged 15 to 25, although the definitions of these phenomena can vary as well as the types of experiences encountered.

b. Definitions of online hate speech and NCII

i. General findings

There are three categories that describe possible definitions used to define online hate speech and NCII among our sample:

- *Description of these behaviours*: This category entails detailing the characteristics and types of NCII or online hate speech. For instance, it may involve *breaking the privacy barrier, face is not needed to be considered intimate*, *way of speaking* or *lack of respect*.
- *Common explanations linked to these behaviours*: This involves exploring the perceived motivations of perpetrators and the understanding of consent. For instance, *by accident, conscious act, depends on education…* This category is more developed in objective 3, related to multiple status and perceived motives.
- *Consequences of these behaviours*: This category addresses the effects and repercussions of the acts of NCII or online hate speech. For instance, *reputational damage, suicide, excluding people, body shaming, break of trust…* This category is examined in objective 2, linked to the perception of harmfulness.

Furthermore, concerning NCII, another category was developed:

- *Responsibilities for NCII* involve the division of responsibility and the characteristics of the relationship and the victim (e.g., *lack of trust beforehand, shared responsibility between victim and perpetrator, love).*

In line with the category related to the description of online hate speech and NCII, some additional results can be shared. First, the team observed variations in the definitions of cyberviolence, NCII, and online hate speech among participants. For example, participants often perceived online hate speech as a subset of cyberbullying. The terms ‘cyberbullying’ and ‘cyberviolence’ were sometimes used interchangeably. This lack of clarity in terminology resulted in a lack of precision regarding the participants' experiential understanding (for more details, see Gangi et al., 2023). Second, both forms of cyberviolence were strongly connected by the participants to fundamental rights. Participants expressed their viewpoints based on the values they deemed most important. A delicate balance between human dignity (e.g., respecting others' freedoms, avoiding harm, not publicly exposing private information) and freedom of expression (e.g., humor, enjoying personal freedoms, sharing content) was evident. Depending on the participants and the context, one fundamental right often took precedence over the other. Third, for NCII and OHS, the theme of consent recurred frequently in participants' responses. The limits of consent and the acceptance of risk were discussed, highlighting the importance of teaching digital natives to identify, express, and respect consent clearly.

ii. Definitions specific to NCII

Our findings have enabled us to identify a category specific to NCII concerning responsibilities. The team can identify the central role of consent for all participants, especially in relation to responsibility. Diverse opinions are present among participants regarding the (non-)responsibility of the victim or subsequent disseminators. The relationship between the victim and the disseminators appears to be a key element, particularly in connection to the reason for sharing and the existing relationship of trust. Moreover, sexting is a familiar territory for adolescents and emerging adults. Underlying motivations for engaging in such behaviour, such as love (Cooper et al., 2016; Glowacz & Goblet, 2020), may also be seen as an additional factor to consider responsibility as shared. Emotions might override rational judgment and forego a more prudent and risk-avoidant approach(e.g., if there was insufficient trust established beforehand).

iii. Definitions specific to OHS

During the analysis of the definitions provided by the sample, two categories of definitions emerged. The first label of definition, which the team have named "aggressive content," is characterised by aggressive expression towards an individual or a representative of a group. The second label, "hateful content," aims at promoting hate towards a representative of a group or an entire group. To go further, the team draws on the understanding of hate as a continuum (Cramer et al., 2022; Schweppe & Perry, 2022). Some authors and organisations describe a pyramid system, starting with biases (stereotypes, remarks, jokes), leading to individual acts of prejudice (harassment, exclusion, dehumanization), which can escalate to discrimination (political, economic, etc.), and ultimately culminate in bias-motivated violence (against individuals or groups) and even genocide (al Serhan & Elareshi, 2019; Anti-Defamation League, 2018).

From a criminological perspective, understanding online hate speech can be aided by the concept of microaggressions (Clark et al., 2011; Constantine, 2007; Sue et al., 2008). Microaggressions encompass intentional and unintentional behaviours and language that minority and/or oppressed individuals experience in their daily lives (Clark et al., 2011; Constantine, 2007; Sue et al., 2008). Sue and colleagues (2008) identify three types of microaggressions: microassaults (intentional), microinsults (unintentional), and microinvalidations (dismissal of lived experiences). Legally, microaggressions may sometimes be considered as such depending on the harm and context, such as in the workplace, but they are often not covered by anti-discrimination laws. Nonetheless, microaggressions contribute to an environment that fosters prejudice (Schweppe & Perry, 2022). Some groups, including the Association of Chiefs of Police (ACPO), advocate for their criminalisation (Schweppe & Perry, 2022). Moreover, concerning adolescents and emerging adults specifically, several authors indicate that tacit approval of certain behaviours can lead to an escalation of language or actions (Hall, 2009; Wieland, 2007).

In the "hateful message" label, participants seem to incorporate the notion of microaggressions, biases, and individual acts of prejudice into their definition of online hate speech when discussing insults, jokes, remarks, and harassment directed at individuals. The "exclusionary message" label may correspond to elements related to discrimination, exclusion, and dehumanisation found in individual acts of prejudice, potentially aligning with a higher level within the pyramid framework. Therefore, comments, remarks, verbal aggression, jokes, insults, and harassment could create a framework conducive to discrimination, exclusion, incitement to violence, and even racism. Simultaneously, the existence of behaviours expressing "exclusionary messages" could authorise and legitimise behaviours related to "hateful messages," supported by a process of normalisation of these behaviours (Costello & Hawdon, 2020).

Furthermore, although not all of these acts may have a discriminatory intent, the focus on identity, combined with their frequency and public nature, can potentially create a sense of exclusion and discrimination, regardless of the perpetrator's intentions.

In addition, individual characteristics have been analysed in-depth for online hate speech (see Gangi & Mathys, in press). To consider the individual characteristics, the team chose to dichotomise these variables, such as sexual orientation (heterosexual, n=13 and non-heterosexual, n=10), cultural background (exclusively of European origin, n=15 and non-European, n=8) and age (15-17 years; n=9 and 18-25 years, n=14). The team found that cultural background exhibited relatively similar trends among young Europeans and non-Europeans, aligning with the two identified labels (aggressive and hateful messages). Among the other individual variables, emerging adults (18-25 years old) of non-heterosexual orientation and female gendertended to formulate definitions that fell into the "aggressive message" label more frequently. In contrast, their male and/or heterosexual counterparts tended to formulate definitions associated with the "hateful message" label.

These findings indicate also that both non-heterosexual individuals and females, as well as older participants (18-25 years old), tend to describe online hate speech with a broader formulation, associated with "aggressive messages." These variables, especially regarding sexual orientation and gender, are frequently subjected to online victimisation (Costello & Hawdon, 2020; Reichelmann et al., 2021). According to Costello and colleagues (2019), online victimised individuals are more likely to define speech as hurtful or harmful based on their past experiences. This victimisation may explain the preference for a broader formulation of online hate speech, encompassing micro-aggressions as well, ultimately reinforcing an amplification of the phenomenon.

Further, these analyses show that male participants and those identifying as heterosexual are more inclined to provide definitions associated with the label of "hateful messages." We have observed in existing literature that these profiles are more closely linked to the role of perpetrators of online hate speech (Bernatzky et al., 2022). In this study, these trends can also be observed, with two out of three perpetrators in the "hateful messages" label being heterosexual males. Regarding cultural background, respondents of non-European origin are similarly distributed between the two categories of formulated definitions, whereas those of European origin predominantly propose definitions related to "aggressive messages." However, when examining the experiences of respondents of non-European origin, the results show that seven out of eight reported being victims and/or perpetrators of online hate speech.

These results may resonate with research on digital platforms, which identifies a significant amount of hate content targeting cultural backgrounds (Costello et al., 2016; Hawdon et al., 2018; Reichelmann et al., 2021). However, unlike previous findings (Costello et al., 2019), the specific experiences of perpetrators and/or victims in this study did not lead to proposing a broader definition of online hate speech. Instead, most of this sub-sample falls into the "aggressive messages" label, both for formulated definitions and experienced incidents. One of the main hypotheses, in addition to the amplification, could be the normalisation of certain behaviours (Costello & Hawdon, 2020). Ortiz's study (2019) explains that individuals from cultural minorities, in the absence of social support from their peers and tired of responding to denigrating content, may turn to desensitization as an adaptation strategy. This mechanism of desensitization may contribute to a broader qualification of what constitutes online hate speech, without placing the focus solely on experienced exclusion.

c. Appreciation of harmfulness

i. General findings

Four similar categories could be linked to the appreciation of harmfulness for online hate speech and NCII among our sample:

- **Criteria related to the audience:** This involves distinguishing between private and public audiences. For instance, *spreading to close contacts (e.g., friends and family) is more hurtful*, *visible for the whole online network is more hurtful*, *private is more hurtful than public*…
- **Criteria related to the actors:** Factors include the characteristics of the perpetrator, such as intent, and the characteristics of the victim. Additionally, the proximity between the perpetrator and the victim plays a role. For instance, *minors involved is more hurtful*, *depending on whether you are in a weak position, depending on the number of perpetrators...*
- **Criteria related to the behaviours:** This category considers the intensity and repetition of the harmful acts. For instance, *whether it is consensual or not*, *words are more hurtful more than physical harm*, *higher amount (of messages/photos) is more hurtful*…
- **Criteria related to the consequences:** This aspect accounts for the impact on both the victims and the perpetrators. For instance, *depends on your background*, *causes discomfort*.

Several of the respondents mentioned the existence of online discussion groups where non-consensual intimate images (NCII) would be prevalent. These groups would communicate on instant messaging applications such as Snapchat or Telegram. The dissemination of NCII in such groups, often called expose groups, impacts on several criteria, namely (i) the criterion of audience, as these groups can sometimes consist of thousands of members, (ii) the criterion of actors, as there are various successive sharers who might not necessarily know the person depicted in the image and (iii) the criterion related to behaviour, since these groups increase the frequency of disseminating; and the criterion related to consequences, as victims discuss challenging repercussions, such as the distress of knowing that their image is in the hands of strangers or acquaintances.

The perception of harmfulness was further developed under the category "virtual/social network". It is composed of two categories:

- **Description** includes the characteristics of the virtual world, and the specificities of the virtual world compared to the real world (*e.g., is not a concrete reality, anonymity, online relations are easier to forget than in real life*)
- **Common beliefs** includes positive and negatives consequences, and perceived risks, and is related to both the social environment and the individual (*e.g., time consuming, facilitating emotional regulation, knowledge*)

This suggests that those participating in our survey are considering the distinct characteristics of the digital environment.

In line with the criteria related to the consequences, it appears that individuals' experiences are influenced by their connection to the digital world. For instance, some people can distance themselves from the digital realm as they perceive it as separate from the real world, while others are profoundly affected by the lasting nature of the online environment. Literature has allowed us to demonstrate the link between exposure to content and the perception of harmfulness of this content (Bernatzky et al., 2022; Costello et al., 2019).

From our results, it appears that in addition to this relationship with the viewed content, the respondents would consider specific digital characteristics (e.g. not a concrete reality, anonymity) to, on one hand, distance themselves from potential negative consequences experienced within the digital realm, or, on the other hand, distance themselves from potential consequences of their own actions towards others.

ii. Appreciation of harmfulness specific to OHS

The findings have revealed four criteria for the perception of harm caused by OHS, all of which are closely related to the virtual/social network category. The team observed certain patterns in participants' responses that highlight the connection between the digital/online realm and identity. Internet is frequently employed to mask, shape, or amplify one's identity. Keipi et al. (2017) introduced the Social Identity model, which posits that in the virtual world, interactions often bolster dynamics of group identity, solidifying specific facets of one's identity. Individuals might come to represent a particular group. Furthermore, online platforms rely on algorithms to tailor content based on users' preferences, which may create an echo chamber that reinforces their identity (Keipi et al., 2017). Theoretical insights and qualitative interviews contribute to an understanding that identity elements (e.g., gender, culture, sexual orientation) are frequently accentuated in the digital realm. Consequently, the sense of group identity diverges from that in the offline world, influencing perceptions of victimisation (e.g., feeling victimised as a bystander due to attacks on one's group) and potentially increasing the perpetration rates (e.g., targeting individuals who do not belong to one's group based on their differing identity characteristics).

iii. Appreciation of harmfulness specific to NCII

From the interviews, it is clear that participants recurrently underscore the victim's responsibility, with a particular focus on the notion of consent for the initial sharing. This finding may impact the harm caused to a victim. Previous research found that the extent of harm of NCII may be contingent on whether the initial image sharing was consensual or not (Dekker et al., 2019). In connection with the results related to definitions, the team also observed that the victim's responsibility is highlighted.

d. Coping mechanisms

Regarding coping, several mechanisms have been identified, such as the use of substances, denial, acceptance, and self-blame. Other mechanisms identified are consistent with literature, including the practice of discussing the incident or requesting the deletion of the images (Wachs et al., 2020). In reaching out for help, it seems that participants often sought support from their social networks, particularly from friends. Some individuals faced challenges when seeking help from institutional sources such as the psychological services of schools, as they reported experiencing rejection. It is worth noting that a significant majority of respondents indicated that they did act in response to their victimisation experiences.

According to the interviews, the participants seem to reject conventional institutions for support and appear to find satisfaction, at least on the surface, in the support of friends. The findings appear to align with the very low figures presented by Wachs and colleagues (2020), stating that only 2% of cyberhate victims report it to the police, and 4% to a professional. The lack of professional support may result in inadequate coping mechanisms, which may explain a sense of isolation, sometimes coupled with guilt. Opting not to share information with family members could result in adverse outcomes. Furthermore, avoidance strategies can be identified, such as creating a new account or changing schools.

e. Status, motives and individual characteristics

i. General findings

It became apparent from the interviews that most of the participants are familiar with OHS and NCII, as they have defined at least one role (witness, perpetrator, and/or victim). Furthermore, it is relevant to note that many of our respondents take on multiple roles, sometimes even being both a bystander, perpetrator, and victim simultaneously. The results are linked to previous results that already highlighted the co-occurrence of status (Costello & Hawdon, 2020). As mentioned, cyberviolence has the unique ability to increase the number of bystanders who may experience a form of victimisation through exposure to certain content. Hence, their perceptions of motivations could have been influenced by their experiences, thereby amplifying the significance of the results of this research. Among the motivations of the perpetrators and those imagined by victims or bystanders, there can be both overlap and differences.

Several of the perceived motivations highlighted by our participants link to scientific literature concerning NCII (Harper et al., 2021; Henry et al., 2019) or online hate speech (Awan, 2014; Jacks & Adler, 2016). Several of the participants suggested intentional motivations, such as revenge, the pursuit of power, or entertainment (Awan, 2014; Harper et al., 2021; Henry et al., 2019; Jacks & Adler, 2016). This type of motivation is directly linked to revenge-porn, a term widely used in the context of NCII, as well as our label "hateful message" for online hate speech. This kind of conceptualisation is not innovative and might even constitute a motivation accepted by the general population. In contrast, this intentional motivation is not the only one, and other motivations are suggested. This therefore implies that our respondents aged 15 to 25 do not merely attribute intentionality to these two behaviours.

As to the motives, four categories of status could be linked to the participants. The perceived motives emerged as a main category (“common explanations”) to discuss the status of our participants involved in online hate speech and NCII:

- Immaturity is related to elements specific to the lack of maturity and the short-term view (*e.g., immature, impulsive, selfish*)
- Intentional motives (*e.g., hurting the other, belittling, manipulation*)
- Emotional motives (*e.g., boredom, jealousy, being hurt*)
- Social motives (*e.g., popularity, showing off “your” trophy, group effect*)

Specific to OHS, another category has emerged:

- Self-expression (*e.g., giving an opinion, persuading, expressing political beliefs)*

This finding leads to further analysis of the identification of status and motivation. First, one of the motivation-related categories is closely linked to immaturity. Immaturity is connected to specific elements, such as only having a short-term perspective, impulsive behaviour, and selfishness. This type of result was not present in existing literature. Therefore, even though this subcategory aligns completely with the identity development of adolescents and emerging adults, it is interesting to establish connections between these elements and other subcategories.

Second, concerning the emotional motives subcategory, it was observed that the respondents expressed hate speech or propagate hate-inciting discourse as a form of personal emotional regulation. Emotional elements such as anger or laughter are present in existing literature (Awan, 2014; Harper et al., 2021; Henry et al., 2019; Jacks & Adler, 2016).

Third, the research showed a link between identity development and the social motive (e.g., fame, group effect, popularity), i.e. a type of motivation that is highly prevalent in the literature for online hate speech and hate discourse (Awan, 2014; Harper et al., 2021; Henry et al., 2019; Jacks & Adler, 2016). This result can be linked to the criteria of harm perception, which often depend on external elements, as is the case with the criteria related to the audience (e.g., *private is more hurtful than public*) and criteria related to the actors (e.g., *depending on the number of perpetrators*).

The sample of this study presents a strong interest in the perception of external individuals and not only a focus on internal factors. This is apparent when looking at the criteria related to the consequences for the victims or the consequences for the perpetrator. These results lead to the hypothesis that participants, due to their age, exhibit specificities that should not be overlooked. Indeed, the process of identity construction typically concludes towards the end of adolescence and the beginning of adulthood (Rocque, 2015). Our sample, with an average age of 19.83 years, still seems to be in a developmental phase, which can have several implications. First, the short-term view may result in impulsive behaviours (immaturity), or behaviours triggered by emotions (emotional motives). Second, the lack of maturity and the limited perspective on situations may lead to selfish decision-making (immaturity) with the objective, notably, of belonging to a group (social motives). Thirdly, the identity reflections that characterise this period might lead individuals to reflect their own feelings on others, whether related to gender, sexual orientation, cultural background, or even intimacy and sexuality.

Regarding these perceived motivations, it has been noted that these explanations may not necessarily align with the actual motivations of offenders. However, these perceived explanations provide insights into how participants understand and interpret the reasons behind online hate speech and NCII. Notably, adolescents and emerging adults found that the perceived motivations for both online hate speech and NCII were generally the same, except for one difference. Hate speech was also seen as a form of self-expression, such as expressing opinions or political views. This perception differs from the legal framework, which primarily focuses on hate speech as a group-targeted behaviour.

ii. Motivations and status specific to OHS

The category related to self-expression (e.g., giving an opinion, persuading, expressing a political opinion) appears to be specific to online hate speech. This finding is inherent to the legal definition of online hate speech and our definition label called 'hateful content'. Thus, while the expression of hate and the goal of exclusion appear to characterise online hate speech (Hawdon et al., 2017), some of our participants do not consider self-expression as the unique form of perceived motives, nor the intentional motives (e.g., anger). Due to the age of our sample (15-25 years old) other types of perceived motives have been indicated by our respondents, such as immaturity and emotional cues.

iii. Motivations and status specific to NCII

In popular literature, NCII is often labelled as ‘revenge porn’, suggesting that revenge is a common motivation for NCII. The diversity of motivations, as reported by the participants in the interviews, encompassing factors such as immaturity, intentional motives, emotional motives, and social motives, demonstrates that adolescents and emerging adults do not solely imply revenge as a motivation. This is supported by existing literature (Harper et al., 2021; Henry et al., 2019).

3.2 REGULATORY FRAMEWORK MAPPING OF OHS AND NCII

3.2.1. METHODOLOGY

This work package aimed to examine how OHS and NCII are tackled on the judicial and legal fronts in Belgium. The objective of this work package is to determine what behaviours, and based on which legal criteria, are considered illegal online content under the current legal framework, doctrine and case law. The project intends to map the several (national and supranational) legal regimes that can be applied in Belgium to these online behaviours, the scope of these legal norms in addressing the several manifestations of such behaviours and the concrete application of the rules in case law.

Step 1: Literature study

The literature study consisted of collecting, reading, and analysing the doctrine related to the definition of OHS and NCII from a legal point of view. The team focused on the Belgian and European legal doctrine, in particular literature concerning the EU approach and the case law of the European Court of Human Rights, as this will directly impact Belgian legislation. Case law has also been taken into consideration, regarding the international literature relevant for understanding the legal framework. Existing literature helped identify the relevant norms for the mapping of the legal framework.

Step 2: Mapping of legal framework

Mapping the legal framework led the team to identify the legal provisions applicable to NCII and OHS at the Belgian level, i.e., national, European supranational and international provisions relevant to NCII and OHS. In this research, both norms explicitly targeting (online) hate speech and NCII were considered, as those are the norms on which courts rely when deciding cases of (online) hate speech and NCII. The relevant norms were further analysed and categorised based on the level of regulation, the binding or non-binding nature, the actor concerned, the tech specific or neutral nature of the norm, and the specific form of NCII or OHS concerned. In addition to mapping the legal framework on OHS and NCII, the research mapped those fundamental rights and values that delineate the limits of criminalising (online) hate speech and NCII, particularly the freedom of expression, right to information and freedom of press. The research showed this analysis to be relevant for OHS.

Step 3: Case study research

The case study research led the team to conduct an extensive analysis of the case law relating to OHS and NCII. There is, in Belgium, little published case law. The team therefore requested authorisation from the Presidents of the criminal Courts of first instance to access the (unpublished) decisions, as well as authorisation from the public prosecutors to be able to read the files that have been dismissed or that have been the subject of an alternative or simplified prosecution procedure. The importance of the task and the lack of cooperation from some judicial authorities led the team to adapt its research strategy. At the level of the courts, the team was granted access to the judgments rendered by the criminal Courts of first instance of Brussels (French-speaking and Dutch-speaking), Liège and Antwerp (Division Antwerp only) relating to OHS and NCII. At the level of the public prosecutor's offices, the team was given access to the relevant files in the judicial districts of Namur, Ghent and Brussels (including Halle-Vilvoorde but only regarding OHS). In order to test the practices specific to the different judicial districts and given the overlap between NCII and OHS on the one hand, and other offences on the other hand, the team sometimes also targeted cases involving harassment (criminal Courts of first instance of Antwerp and Liège) and dissemination of child pornography (criminal Court of first instance of Brussels, public prosecutor office in Namur). After manual filtering, the team retained a total of 193 files concerning OHS and 423 files concerning NCII.

The objective of the analysis was to have an overall view and to see the possible distinction between cases brought before the Courts and those that remained at the stage of the public prosecutor's office, because they had been dismissed or had been the subject of an alternative or simplified prosecution procedure, such as a settlement or a mediation.

The methodology used for the case study research is the coding technique (Lawless et al. 2016). This technique allows the team to pre-select different criteria to retain several features of the case and then be able to identify trends. The selected criteria can be grouped into different categories related to the context of the case, the perpetrator and the victim (gender, age, sexual orientation, nationality, relationship with the perpetrator,…), the possible other parties involved in the case (OSPs, specific organisations like Unia, Child Focus, the Institute for equality of Women and Men,…), the source of the case, the possible other offence(s) prosecuted at the same time, and the outcome of the case. There were some variations between the criteria the team used for cases involving NCII and those for hate speech according to the specific characteristics of both, but the main categories remained the same. The codes were systematically applied to all cases by the same group of coders.

Step 4. Analysis and hypothesising

In the final step, the team compared the outcomes from the legal mapping with the case study research to better understand the delineation between lawful and unlawful OHS and NCII as well as the reasons for the limited number of actual cases in Belgian courts. It was apparent from the coding that some criteria could not be systematically coded since the cases did not all provide the necessary information, in consequence of which these data were not taken up in the analysis. Based on the analysis, the team developed hypotheses for further research, conclusions and recommendations.

3.2.2. RESULTS

a. The Legal Framework of OHS

i. Mapping the legal framework of OHS

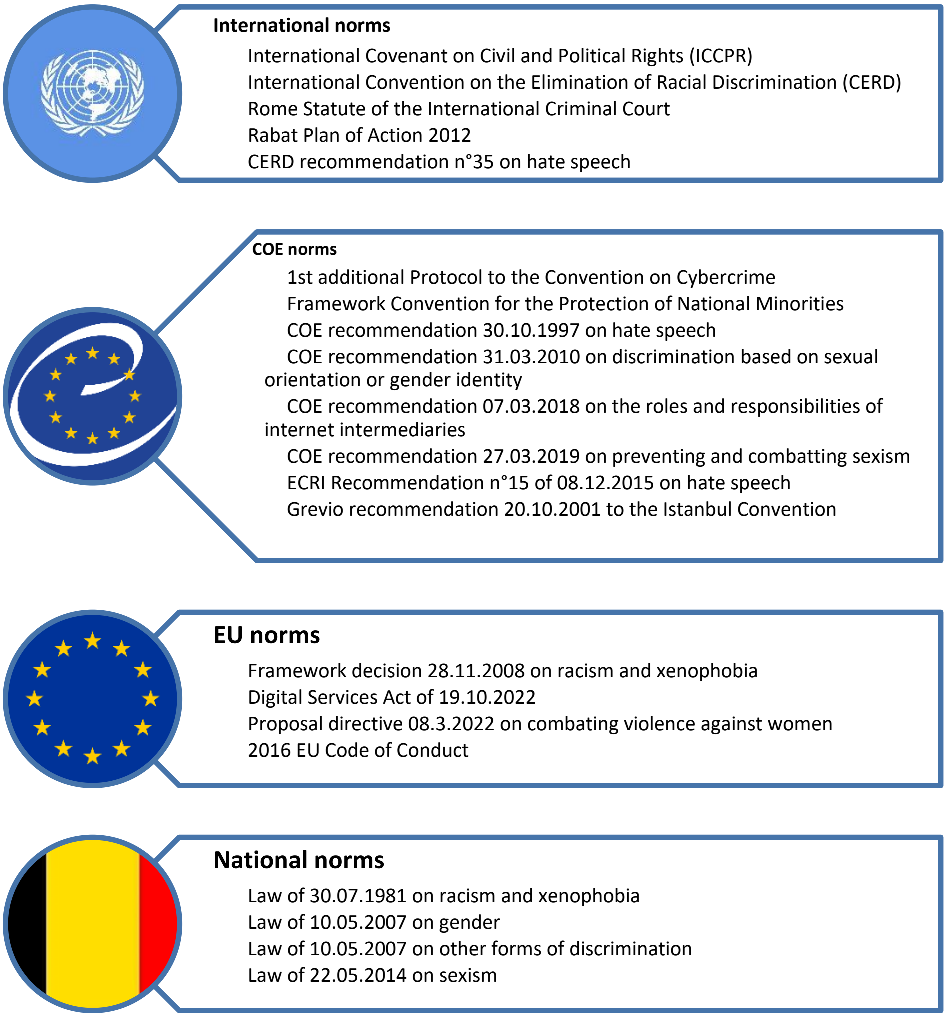
The first part of the research on the legal framework of OHS intended to map the full legal framework of norms at the disposal of the Belgian national jurisdiction to address cases of hate speech. The purpose of this mapping exercise is to understand the adjudication of online speech targeting persons or groups on their status or personal characteristics as lawful or unlawful speech.

In order to map this framework, the research's purpose was to delineate what constitutes norms addressing 'hate speech'. There is no commonly accepted definition in national, European or international law of what constitutes 'hate speech' and by extension, what constitutes 'OHS'. Defining hate speech was previously described as 'a seemingly elusive task' (Fino, 2022). There are, however, several regulations at the national, European and international level that address hate speech explicitly, namely norms that indicate what forms of speech are unlawful because it targets a person or group of persons for their personal characteristics or status, or norms that have been relied on by national courts when addressing whether hate speech is to be considered unlawful. At the same time, there are legal norms at the national, European supranational and international level that limit the power of authorities to criminalise hate speech, i.e., in the light of other rights – in particular the freedom of expression and right to information – and values. As such, the legal framework of hate speech is a layered structure with rules both forming and limiting the delineation of what constitutes lawful and unlawful hate speech.

The mapping exercise showed that there are 4 levels of regulation that are relevant for delineating whether alleged hate speech is to be considered lawful or unlawful speech (the full mapping is to be found in annex 7).

- **Level 1: international, supranational and national norms explicitly addressing hate speech**

The research mapped the several legal norms that explicitly address hate speech to discern the framework for the mapping exercise, which focused first on those legal norms at the international, European supranational (EU and COE) and national level that explicitly consider hate speech. The following norms were analysed:



International norms

International Covenant on Civil and Political Rights (ICCPR)
International Convention on the Elimination of Racial Discrimination (CERD)
Rome Statute of the International Criminal Court
Rabat Plan of Action 2012
CERD recommendation n°35 on hate speech

COE norms

1st additional Protocol to the Convention on Cybercrime
Framework Convention for the Protection of National Minorities
COE recommendation 30.10.1997 on hate speech
COE recommendation 31.03.2010 on discrimination based on sexual orientation or gender identity
COE recommendation 07.03.2018 on the roles and responsibilities of internet intermediaries
COE recommendation 27.03.2019 on preventing and combatting sexism
ECRI Recommendation n°15 of 08.12.2015 on hate speech
Grevio recommendation 20.10.2001 to the Istanbul Convention

EU norms

Framework decision 28.11.2008 on racism and xenophobia
Digital Services Act of 19.10.2022
Proposal directive 08.3.2022 on combating violence against women
2016 EU Code of Conduct

National norms

Law of 30.07.1981 on racism and xenophobia
Law of 10.05.2007 on gender
Law of 10.05.2007 on other forms of discrimination
Law of 22.05.2014 on sexism

Figure 3. Norms on OHS for mapping the legal framework

In the mapping exercise, norms that address speech on other bases than inciting violence, discrimination, segregation or hate against a certain person or group on the basis of a personal characteristic or status, were excluded. Examples of these include norms criminalising the glorification of terrorism or negationist legislation criminalising the denial or minimalization of certain historic genocidal acts, even though there is often an important overlap with those acts and OHS on the basis of religion, race or nationality. e.g., Holocaust denial and OHS against Jews and therefore, considered hate speech (*ECtHR 15 October 2015, Perinçek v. Switzerland, § 253*).

The several legal norms applicable to OHS can be categorised based on the following criteria.

Level of regulation	National, supranational (EU or Council of Europe), international
Binding	Binding norms / proposal for binding norms on Belgium (signed and ratified) or non-binding (soft) law e.g., guidelines, …
Actor	Rules regarding unlawful speech binding on state only (state obligations), natural and legal persons (e.g., criminalising certain behaviour) or other specific other actors (e.g., industry such as social media)
Criterion	Focused on all forms of hate speech (general) or only on one or more specific grounds for hate speech (e.g., gender)
Technology	Focus on or special consideration for online unlawful speech (tech-specific) or on all speech including online speech (neutral)
Content	Unlawfulness of speech based on the hateful content of the speech (content) or on a call to action in the speech (incitement)
Intent	Unlawfulness of speech (not) depending on certain intent e.g., to cause harm

Table II. Categorisation criteria for OHS norms

The in-depth analysis of the key legal norms on hate speech resulted in several findings:

1) **Level of regulation**

At all levels, binding norms can be found addressing hate speech. In general, the norms at the international and supranational level are drafted as positive obligations on states to tackle hate speech and on the national level as prohibitions vis-à-vis natural and legal persons to engage in hate speech. The vast number of norms at the international and supranational European level require the state to criminalise (online) hate speech and provide effective remedies to victims. However, exceptionally, norms at the international or supranational level also address natural and/or legal persons directly, e.g., international criminal law rules regarding incitement to genocide or EU rules addressing online digital services.

In some norms, the positive obligations on the state to combat hate speech is not limited to criminalising (online) hate speech but extends to other measures for less serious forms of hate speech. For example, the COE 2022 Recommendation on hate speech distinguishes between (i) hate speech that is prohibited under criminal law , (ii) hate speech that does not attain the level of severity required for criminal liability but is subject to civil or administrative law , and (iii) 'offensive or harmful types of expression' which are not sufficiently severe to be legitimately restricted in view of the rights and freedoms entrenched in the ECHR, but call for alternative responses such as counterspeech (Recommendation CM/Rec(2022)16[1] of the Committee of Ministers to member States on combating hate speech, Adopted by the Committee of Ministers on 20 May 2022 at the 132nd Session of the Committee of Ministers). However, at the national level, the tendency in norms is clearly to opt for criminalisation for unlawful hate speech.

2) **Binding vs non-binding**

The analysis shows that at the national, supranational European – in particular EU, and international level there is already a framework of norms addressing hate speech. This includes binding norms at all levels.

3) **Actors**

The legal framework is composed of on the one hand international and European supranational norms holding positive obligations on the state to act against hate speech, and on the other hand international, European and national norms that are directly applicable to natural and/or legal persons. The international and European norms either hold explicit obligations on states to criminalise hate speech (e.g., art. 3 of the COE additional protocol to the Cybercrime Convention requires states to 'establish as criminal offences distributing, or otherwise making available' racist and xenophobic material to the public through a computer system) or broader obligations to take legal actions and measures without explicitly requiring criminalisation (e.g., art. 20 International Covenant on Civil and Political Rights requires states to 'prohibit' advocacy of national, racial or religious hatred).

More recent norms also address online platforms and search engines, or broader 'IT companies'. It suggests that there is a growing understanding that regulating these companies is vital for tackling OHS. Such norms, binding or non-binding, are mostly found at the European supranational level. This can be explained by the intrinsically cross-border nature of hosting services where OHS occurs.

4) **Criterion**

Whereas at the national level there already exists an all-encompassing framework on hate speech covering hate speech on a vast array of potential individual characteristics, at the European supranational and international level the protection of hate speech is limited to a small number of grounds. At the supranational and international level, regulation of hate speech mostly focuses on race, skin colour, nationality, ethnicity and religion. In recent years, there is a growing body of norms focusing on hate speech based on gender, whether labelled as sexism or GBV. As such, there exist generic norms on hate speech addressing hate speech irrespective of the specific targeted personal characteristic or status of the victim(s) as well as specific norms on hate speech addressing hate speech targeting victim(s) based on a specific ground, e.g., race or gender. Consequently, there is a fragmented legal framework based on the specific characteristic or ground of the victim(s).

5) **Tech**

Almost all norms are technology-neutral, focusing on hate speech as such rather than focusing on cyberviolence or OHS. An 'online'-specific approach is rather found in non-binding documents. A notable exception is the first additional protocol to the Cybercrime Convention that focuses specifically on online hate crimes (First Additional Protocol to the Convention on Cybercrime, concerning the criminalisation of acts of a racist and xenophobic nature committed through computer systems (ETS No. 189)). This binding Convention, extending the Convention on Cybercrime's scope, was, however, not ratified by Belgium.

6) **Focus**

Binding legal norms – at the national, European supranational and international level – generally reserve limitations of speech, in particular criminalisation, for action-oriented hate speech, e.g., inciting to hatred, discrimination or violence. European and international –particularly non-binding - documents also problematise content focused hate speech, i.e., speech without call to action, but do not call for criminalisation of such behaviour.

As such, within the broader category of 'hate speech' there appears a core, i.e., action-oriented hate speech, that is considered particularly problematic. The COE's European Commission against Racism and Intolerance recommendation on hate speech states, for example, that *'the use of hate speech may be intended to incite, or reasonably expected to have the effect of inciting others to commit, acts of violence, intimidation, hostility or discrimination against those who are targeted by it [...] is an especially serious form of such speech*' (ECRI General Policy Recommendation no. 15, CRI(2016)15). However, criminalisation of content-focused norms is more prominent when addressing sexism, i.e., gender-based hate speech.

7) **Intent**

The vast number of norms do not take into account the intent or purpose of the person disseminating hate speech. Such intent or purpose, however, can be found in national norms criminalising hate speech. While not provided in the provisions of the anti-discrimination and racism legislations, the Belgian Constitutional Court argued that there needs to be 'a particular, malicious will to incite' discrimination, hate, violence or segregation (Belgian Constitutional Court October 6, 2004, n° 157/2004). Also, in the Belgian national anti-sexism legislation, a particular intent is provided, namely that the perpetrator was to 'express contempt towards a person because of his gender, or to regard him, for the same reason, as inferior or to reduce to its gender dimension', as well as a threshold of harm, namely that the behaviour should 'result in a serious impairment of that person's dignity' (Law of 22 May 2014 on combating sexism in public spaces and amending the Law of 10 May 2007 on combating discrimination in order to punish the act of discrimination).

- **Level 2: added layer of norms that do not explicitly mention hate speech**

Secondly, the mapping exercise focused on those legal norms that do not explicitly mention hate speech but are considered to intrinsically address cases of hate speech. These norms further enrich the legal framework on hate speech. Literature and the court cases within the coding study were analysed to further map the second layer of the legal framework on hate speech. The research found two such instances of added layers on hate speech.

First, courts, public organisations or scholarly research may read positive obligations on the state to act against hate speech in norms concerning non-discrimination and/or equality clauses, and the protection of physical, mental, and psychological integrity. The mapping study found several binding norms in which positive obligations to tackle hate speech were read:

- The clearest example is the European Court of Human Rights, which reads a positive obligation to protect individuals or a community against stereotyping, stirring up prejudice, incitement to hatred, discrimination and violence on the ground of their status or characteristics (ECtHR 14 Januari 2020, *Beizaras and Levickas v. Lithuania*; ECtHR 16 February 2021, *Behar and Gutman v. Bulgaria*; ECtHR 16 February 2021, *Budinova and Chaprazov v. Bulgaria*).
- Recommendations and comments by expert bodies to the Convention on the Elimination of All Forms of Discrimination against Women (CEDAW) as the COE's Istanbul Convention on Preventing and Combating Violence Against Women and Domestic Violence (Istanbul Convention) read a positive obligation on states to take measures against (online) speech that incites discrimination or violence against women, e.g., the CEDAW Committee recommended states in General Comment n°35 on the elimination of violence against women to adopt and implement measures to encourage media, including in online and other digital environments, to eliminate discrimination against women, including harmful stereotyping (General recommendation No. 35 (2017) on gender-based violence against women, updating general recommendation No. 19 (1992) CEDAW/C/GC/35).

National courts confronted with cases of hate speech have also relied on other incriminations of criminal law that do not specifically target hate speech but have a general application irrespective of the status or characteristics of the victim. This approach is particularly found in the case of OHS. For example, certain forms of hateful speech can be equated to cyber harassment, which is included in an article in the Belgian Act on Electronic Communication (article 145 §3bis Law of 13 June 2005 on Electronic Communication). Other examples include relying on provisions concerning threatening a person with an attack (article 327 Criminal Code) or harassment (article 442bis Criminal Code).

- **Level 3: framing level of human rights protection limiting actions against hate speech**

The third level of the legal framework constitutes the limitations of actions against hate speech, i.e., those forms of speech that may target a person or group based on their status or characteristics but are nevertheless considered acceptable. This framework is formed by the national constitutional and supranational/international human rights protection of the freedom of expression, right to information and freedom of press. The mapping examined whether the binding international and supranational documents holding positive obligations or incriminations on hate speech also hold a reference to the freedom of expression, right to information and freedom of press directly or indirectly by reference to the protection of human rights.

Such reference can be either directly, i.e., whereby the protection for the freedom of expression is considered as a limitation for tackling hate speech, or as a separate article, whereby the protection against hate speech is to be balanced vis-à-vis the freedom of expression. In the first case, the freedom of expression is construed as a limitation to the criminalisation of OHS, e.g., in the first additional protocol to the Cybercrime Convention, the third paragraph provides that member states can reserve the right not to criminalise certain forms of OHS in 'those cases of discrimination for which, due to established principles in its national legal system concerning freedom of expression, it cannot provide for effective remedies'. In the second case, the positive obligation and the protection of freedom of expression are to be balanced.

The analysis shows that most international and supranational documents that include a positive obligation to tackle hate speech, refer to freedom of expression as a limitation of actions and measures to address hate speech, e.g., when criminalising certain forms of hate speech. In these documents, freedom of expression is an implicit limitation in tackling hate speech. In other documents, in particular general human rights law treaties that are not specific to hate speech, there is no specific reference to the freedom of expression in those provisions holding positive obligations to tackle hate speech (i.e., article 20 ICCPR and article 8 joint 14 ECHR). However, in reverse, the protection of the rights of others is included as a limitation to the freedom of expression. This means that in concrete cases of hate speech, courts will have to balance conflicting rights: freedom of expression and right to information on the one hand and the protection of equality, non-discrimination, and the protection of personal integrity on the other.

In contrast, no reference is made to freedom of expression in the international or national norms describing the criminalisation of natural and legal persons for hate speech. That does not mean that these norms criminalising hate speech will not be tested in the light of the freedom of expression, right to information and freedom of press. There is already an impressive body of literature studying to what extent authorities can limit the freedom of expression by criminalising OHS (Buyse, 2014; Racolța & Verteș-Olteanu, 2019; Mchangama & Alkiviadou, 2021). Criminal provisions will need to comply with constitutional, supranational or international protection of freedom of speech. Previous research shows that at the supranational European level as well as at the Belgian national level, the key criteria for balancing freedom of expression and tackling hate speech are based on the case law of the European Court of Human Rights (Petersen, 2021; Van de Heyning, 2022). The analysis shows that the national courts, including the highest courts, implement the case law of the ECtHR rather than developing their own test based on the constitutional protection of freedom of speech. As such, the ECtHR case law on the balancing between (criminalising) hate speech and freedom of expression sets the standard.

In this regard, the research further scrutinised the role of digitalisation in this balance between freedom of expression and addressing hate speech. The mapping exercise quickly established that human rights documents and constitutional rights holding the protection of the freedom of expression in general do not distinguish between online and offline expression. However, certain public authorities and courts are reflective of the impact of digitalisation when considering limitations to the freedom of expression to tackle hate speech:

- The UN Rabat plan holds that the reach, speed and frequency of hate speech are to be taken into consideration and therefore, the mode of communication, with explicit reference to 'internet', is relevant to decide whether speech can be limited to tackle hate speech (Rabat Plan of Action on the prohibition of advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence (A/HRC/22/17/Add.4));
- The ECtHR held that the use of internet (and social media) for disseminating hate speech is relevant for adjudicating whether the criminalisation of such speech is a proportionate limitation of the freedom of expression because of the speed of dissemination, the fact that vast amounts of data can be stored, the fact that content cannot easily be removed, and the fact that content can be disseminated cross-border (ECtHR 16 June 2015, *Delfi AS v. Estonia*, § 133-134; ECtHR 7 February 2017, *Phil v. Sweden*; ECtHR 28 August 2018, *Terentyev v. Russia*, § 80; ECtHR 15 May 2023, *Sanchez v. France*).

- **Level 4: fundamental principles for shaping the balance of hate speech and freedom of expression**

In the mapping exercise of the documents and case law on hate speech, a final added layer is apparent for shaping the legal framework on hate speech. This final layer references to the fundamental principles that are decisive for the balancing between freedom of expression and right to information on the one hand and tackling hate speech from a perspective of non-discrimination, equality, and protection of personal integrity on the other hand.

A clear example can be found in the case law of the European Court of Human Rights, where democracy is put forward as a key principle to define the line between tackling unlawful hate speech and freedom of expression. The ECtHR argues that the principle of democracy includes tolerance, diversity, and pluralism. This principle on the one hand supports caution when limiting freedom of expression, because a democracy thrives on a plurality of opinions. As such, the ECtHR repeatedly states that a democratic state needs to accept opinions that 'offend, shock or disturb' (ECtHR 7 December 1976, *Handyside v. the United Kingdom*). On the other hand, states have a margin to limit the freedom of expression in order to tackle speech that incites violence or discrimination of individuals or groups on their status or personal characteristics, as such speech undermines tolerance and might silence certain minorities. As these principles work both in favour of tackling hate speech as in favour of allowing a broad protection of speech, the states are to develop a framework that delicately balances two strains of rights, i.e., the freedom of expression and right to information on the one hand and the protection of equality, non-discrimination, and protection of physical and mental integrity on the other.

The mapping exercise unearthed a set of principles that are referred to as guiding the balancing of tackling hate speech and freedom of expression. The mapping exercise shows that all international and supranational treaties addressing hate speech include fundamental values to guide the balancing between the freedom of expression and right to information on the one hand and tackling hate speech on the other hand. The most common values in this respect are freedom, democracy and equality, whereby values of tolerance, pluralism, diversity and human dignity are often seen as essential elements of a democratic and equal society. In the mapping exercise, security is a less common value, but particularly prominent in those documents on gendered hate speech. However, the principle of the rule of law is not mentioned in those documents focused on gender, but more common in norms focusing on racism and xenophobia.

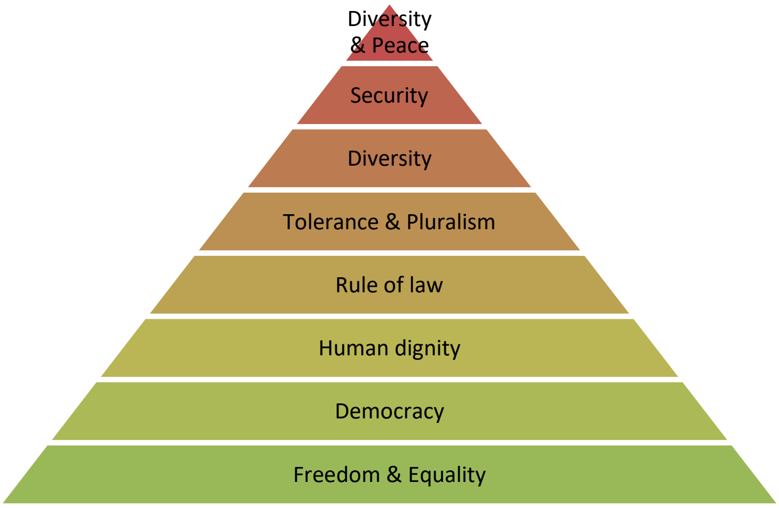


Figure 4. Number of citations in norms on OHS (the most recited principles at the bottom to the least referenced values at the top).

No such values are included in those documents or provisions relevant to criminalising hate speech. As such, when courts are to decide whether certain speech constitutes unlawful hate speech or should be tolerated in the light of the freedom of expression, the values incremental to the supranational and international level are guiding the courts in balancing the respective rights and interests at play.

ii. Coding of criminal cases on hate speech

As was clear from the mapping exercise, OHS can be prosecuted based on several norms criminalising hate speech on specific grounds. In order to assess how OHS is prosecuted and examined, the team focused on all cases which were categorised by the public prosecution as hate speech offences. The team got access to all hate speech cases at four prosecution offices (East Flanders, Halle-Vilvoorde, Brussels, and Namur) in the period 2018 - 2021 and at three courts (Liège, Brussels, and Antwerp) in the period 2016-2021. For reasons of feasibility as well as based on access to the relevant cases, the team selected prosecution offices and courts evenly distributed between the different parts of the country. Given the sufficiently broad selection, dissemination across different public prosecutors' offices and courts, and sufficiently large number of cases, the analysis provides a representative picture of the totality of complaints about OHS in Belgium.

The body of 1562 selected cases were further filtered with a focus on OHS. Of this total body, 193 cases dealing with OHS were retained. These cases were coded based on several criteria relevant to the speech under review, the characteristics of the victim and suspect, the social media or other digital means used and the outcome of the case. From the filtering exercise, the team learned that it is time consuming and difficult to filter out relevant cases within the databases of the public ministry and the court. There is no separate categorisation for OHS cases. This means that not only for researchers, but also for judges and public prosecutors, it is difficult to get a good overview of similar cases and case law on this point (as previously remarked by Commissie voor de evaluatie van de federale antidiscriminatiewetten, 2022).

- **The number of complaints and cases**

The study found that complaints about OHS are exceptional. Only 193 of the 1562 cases investigated on racism, xenophobia and other discrimination were found to be about OHS. These cases dealt with both public hate speech on public websites or social media and targeted hate speech in messages to an individual. Most of the cases started after a complaint by the victim or relative. Exceptionally, complaints were filed by UNIA or bystanders, or started by law enforcement itself, based on public information or information provided by other public authorities, such as intelligence services.

The limited number of actual complaints is remarkable, given that 1) the mapping exercise showed a broad criminalisation of hate speech also applicable to OHS, and 2) the prevalence research as included in 3.3 of this report shows that the population mentioned to very frequently see OHS and a substantial number to fall victim of hate speech. The latter figures show that OHS is very common online. Yet, this does not translate to actual complaints.

- **Legal basis for complaints**

The vast majority of complaints was filed on the basis of the anti-racism law (144 cases). Other legal basis for the complaints were the acts on negationism (law of 23.03.1995), sexual orientation (law of 10.05.2007 on sexual orientation), gender (law of 10.05.2007 on gender), disability (law of 10.05.2007 on disability) or other protected criteria (law of 10.05.2007 on other forms of discrimination). Further, complaints had been filed on (cyber)harassment.

- **Discontinuation of prosecution**

From this number of 193 cases, the vast majority of complaints filed were discontinued. The prosecution dismissed the cases on a wide variety of reasons, namely lack of evidence, exceeding reasonable time for the prosecution, other priorities, lack of capacity, unknown perpetrator, the suspect being a first offender, the assessment that no offense had occurred, and the disproportionality between facts and potential prosecution. In a number of cases, the complaint was dismissed because the situation was regularised, e.g., because the suspect deleted the post or the suspect apologised to the victim.

Of the 193 cases selected, only 30 cases ended up in court, which is an absolute minority. The vast number of cases were discontinued by the public prosecution. In a limited number of cases, the prosecution handled the cases alternatively, e.g., by means of mediation or probation. In the stock of cases, the team found several comparable cases with known suspects, i.e., cases with quasi-identical or comparable relevant facts, whereby in some cases a choice was made for prosecution and in other cases there was a discontinuation of the case. The team also found a significant stock of complaints that were clearly lawful expressions under Belgian criminal provisions on hate speech.

It shows that victims cannot easily navigate what forms of OHS are unlawful or lawful. The limited number of judgments, the limited publication of such judgments and difficulty to navigate the database to find cases, are elements that further decrease the foreseeability and transparency for victims and other actors to delineate lawful from unlawful OHS.

- **The grounds for alleged unlawful hate speech**

The vast majority of the cases concern hate speech against people of a different nationality (xenophobia), hatred based on skin colour and race, and hatred based on religion. The most prominent grounds for OHS were a black skin colour and/or being of African origin (33 complaints), a tanned skin colour and/or being of North African or Turkish origin (24 complaints), or religion, particularly Muslims (43 complaints) and Jews (33 complaints). Not only nationality itself, but also nationality status was found to play an important role in hate speech, namely hate speech against migrants or foreigners (21 complaints). The limited 'residual category' (37 complaints) included cases where there was hate speech based on sexual orientation and gender (identity), with a focus on homosexuals (16 complaints) and women (12 complaints).

The research found that in many of the cases several of the grounds were invoked or several of these grounds were touched upon in the alleged hateful speech, e.g., threats both on the basis of the religion of the person and the skin colour. In that respect, the research found several recurrent 'clusters of hate' whereby the same grounds were targeted by a suspect in one comment or post. The research could discern 3 such clusters (see Figure 5).

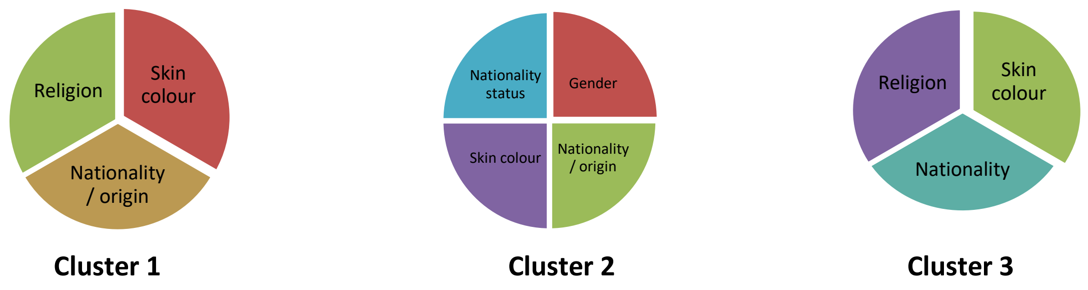


Figure 5. Clusters of hate

- **Means of dissemination of alleged unlawful hate speech**

The research found a wide variety of the means of expression of the alleged hate speech, namely sending a direct message or email, posting a remark, a picture, meme or livestream on social media, forwarding or posting a link to a racist post, hyperlinking an article with allegedly hateful content, posting a response under another social media post, or posting a hateful comment on an online news site or platform. In the stock, one case dealt with impersonation, whereby the suspect created a fake account in the name of the victim, posting racist comments on this fake account. Another case focused on the lack of action taken by an administrator when several racist comments were posted on his Facebook wall. The stock included a comparable sample of complaints concerning public posts and direct interactions between the suspect and the victim without publicity, e.g., by sending emails, texts or DM's.

Notwithstanding this wide variety of means of expression, in most cases (165 cases) the complaint concerned the alleged unlawful content of a text of the post or message. Only in a limited number of cases, the complaint only targeted the image or video (livestream or uploaded video). Images posted alongside text appeared to be considered rather supporting the hateful content than the focus of the complaint itself. The research further remarked that in the stock of cases a variety of expressions of OHS can be discerned, namely insults and derogatory remarks, threats, and incitements to violence, discrimination, hatred, or segregation. The most common form of expressions in the stock were insults with the use of swear words and derogatory naming (e.g., 'apes' or 'whores').

Facebook appeared the most popular means for disseminating the alleged unlawful hate speech (110 cases). With 12 complaints focusing on alleged unlawful hate speech via Messenger and 6 via Instagram, the Meta-group applications are particularly represented in the stock of cases. The second, but incomparably less represented, social medium featuring in the complaints is Twitter (13 complaints).

- **The perpetrators and victims**

Of the known suspects, 36 were female and 116 were male. As to the victims, there is no significant difference based on gender, with 45 of the known victims to be female and 47 of the known victims to be male. Both as to victims as suspects, there is an even dissemination among all age categories. The victims are predominantly from a different ethnic or cultural background or hold a different nationality. This finding goes together with the fact that the vast majority of complaints concern racism and xenophobia. Where the complaints concern alleged unlawful hate speech via direct messages, there is a higher number of victims and/or suspects who indicate to know each other, e.g., colleagues, neighbours or former lovers and friends.

A substantial part of the suspects denied having posted the content or states not to remember having posted the comment or sent the message. As to the motives for posting, three reasons are prevalent: 1) for fun, joke or satire, 2) in reaction to a concrete situation in the suspect's life (e.g., revenge, frustration, bullying, …), 3) reaction to perceived injustice or silencing. Suspects who argue that they posted the comment as a reaction to injustice or silencing in society, particularly make reference to freedom of expression or explicitly deny the unlawful nature of the comment.

iii. Analysis of case law research in the light of the legal framework mapping

The findings of the case law study are surprising in the light of the legal framework mapping study. The latter found that there exists an all-encompassing framework for sanctioning OHS and this based on legal norms sanctioning incitement to discrimination, violence, hatred, or segregation based on a non-exhaustive list of personal status and characteristics. Moreover, this study found that Belgium is under several overlapping supranational and international obligations to tackle (online) hate speech. Nevertheless, the case law study found that 1) only few complaints are filed for OHS and 2) these few complaints are only exceptionally prosecuted before criminal courts as the vast part of complaints is dismissed, even though there is a known suspect.

This finding also has an impact on the development of a Belgian framework on the delineation between lawful and unlawful hateful speech. First, the limited number of cases leads to a lack of a developed body of jurisprudence that may guide future victims or bystanders to assess whether a complaint may be successful or for the broader population to assess whether their online behaviour is lawful.

Further, the very limited number of actual cases on OHS before the courts leads to little jurisprudence to develop a detailed testing on the delineation between unlawful hate speech and lawful (to be tolerated) hateful speech in the light of the balancing between the freedom of expression, right to information and freedom of press on the one hand and the tackling of hate speech from the perspective of the protection of equality, non-discrimination and the protection of personal integrity on the other. This explains why there is little constitutional jurisprudential development on this balancing and the ECtHR case law sets the key criteria in this respect.

The research developed several hypotheses to explain the low number of complaints and the low number of actually prosecuted cases.

- **Hypothesis 1: limited incidence of OHS**

A first hypothesis could be that OHS does not occur as often as expected, resulting in only a limited number of cases. This hypothesis could easily be dismissed by the results from WP3 as well as the literature study that show a high prevalence of OHS as reported by respondents. Relevant for the number of complaints is not so much the actual prevalence of OHS, but the number of persons that consider themselves or their relatives victims of such speech. This is clear from the case law study as the vast majority of the cases originated from complaints made by victims and their relatives at the police station.

- **Hypothesis 2: low trust in law enforcement's investigatory capacities**

A second hypothesis is that there is a low readiness to file a complaint for incidents of OHS. A broader claim could be made that victims of discrimination and hate speech in general are less willing to file a complaint, as they often belong to minorities and/or marginalised groups with negative or distrusting tendencies towards law enforcement. However, this does not explain the low number of complaints for OHS in itself, as the case study research at the initial selection and filtering level showed that only a limited number of the overall sample of hate speech cases at the public prosecution and courts consisted of OHS cases.

Another explanation for the low readiness to file a complaint for OHS may be found in the victims' and bystanders' assessment of the lack of effectiveness of law enforcement and justice when dealing with digital crimes and accompanying phenomena. Previous research on cybercrime indicates that victims of digital crimes are less inclined to file a complaint with the police as they assume that law enforcement is not well equipped, acquainted, and knowledgeable of the digital world and, therefore, filing a complaint will not result in effective investigation and prosecution (Van de Weijer, 2020). The case study analysis will support rather than refute this assessment, as the research shows that the vast majority of complaints are dismissed for a number of reasons. It has been previously remarked that the specialised police services are understaffed for tackling online hate speech (Commissie voor de evaluatie van de federale antidiscriminatiewetten, 2022).

Yet, when zooming in on the reasons for dismissal, digital skills are not the predominant reason. Only in 9 cases was the complaint dismissed for insufficient capacity, in 15 cases because the perpetrator was unknown. As such, there might be a gap between the perception and the actual skills of law enforcement to investigate such cases.

- **Hypothesis 3: low expectancy of effective prosecution**

This brings the team to the third hypothesis, namely that victims and bystanders assess that filing a complaint will not result in prosecution or other consequences for the perpetrators. It is indeed remarkable that notwithstanding the all-encompassing framework in Belgium for criminalising OHS, only a few cases are being prosecuted.

The reason can in the first place be found in the procedural framework for prosecuting OHS. Article 150 of the Belgian Constitution provides that press crimes are to be prosecuted for the Court of Assize, which is an ad hoc jury tribunal. The procedure leading up to such proceedings is considered burdensome and time-consuming by prosecution. Belgian highest courts interpreted a press crime as the public written expression of an opinion, contrary to criminal law. The historic origins of this constitutional provision are to be found in the protection of the traditional press against censorship. However, it was argued by the highest courts that a written opinion that is publicly available on internet and contrary to criminal law, constitutes a press crime. As such, victims and bystanders might assess that filing a complaint for such public posts is of little purpose, because law enforcement will be unwilling to bring such cases before the Court of Assize.

This hypothesis of pre-filtering might be supported by the case study analysis that shows that the vast majority of complaints concern those on race, ethnicity and nationality. Article 150 Constitution exempts the antiracism law from its scope. As such, complaints based on the in this act enumerated grounds can be prosecuted before criminal courts. In any case, based on the case study research and literature study, it is clear that article 150 Constitution provides a high threshold to prosecute cases of OHS on public platforms. This is problematic for three reasons:

1) An unsatisfactory situation is created where Belgian criminal law provides an all-encompassing criminalisation of hate speech but in practice does not prosecute such cases;
2) The exemption on the antiracism law creates a certain ‘hierarchy of evils’ in prosecution and is at odds with the reality, whereby rather than hate speech on the basis of one ground, the research finds that OHS often translates to clusters of hate where a person or group is attacked on the basis of several characteristics or status;
3) Because of the threshold, Belgium is at odds with its European supranational and international obligations to require effective action against OHS in general and with the EU Framework on racism and xenophobia as well as the additional protocol to the Cybercrime Convention that also includes religion as one of the grounds on the basis of which effective criminal prosecution needs to be available.

However, the threshold of article 150 Constitution only explains a part of the low level of prosecution. As the case law study shows, an important part of the cases concern one-to-one messages of OHS. In such cases, the threshold of article 150 Constitution does not apply, as the required publicity of the message to constitute a press offense is not met. Also, these cases are dismissed for a variety of reasons, ranging from finding that there is insufficient evidence, over non-priority of the case, to finding prosecution disproportionate to the facts. Furthermore, many of the complaints are not unlawful hate speech under the Belgian rules governing OHS. Analysing the case study, two reasons come to the fore:

1) In many cases, the prosecution finds a lack of evidence because the criminal intent of ‘malicious’ incitement is not obvious. Many suspects mentioned other reasons for posting the OHS, such as for fun, to highlight an injustice or out of anger for the behaviour of the counterparty.

2) In several cases, the prosecution argued that the required 'publicity' of the OHS was not there and therefore the incitement to violence does not constitute a crime. However, when analysing the case, it is clear that several cases could have been prosecuted on other grounds e.g., harassment with a discriminatory intent, cyberharassment or threats, which might result in the conclusion that either there is a lack of knowledge on the legal framework, a lack of willingness, or a lack of prioritisation. This is further supported by the fact that similar cases in the case study were dismissed for different reasons, which shows a lack of coherence and foreseeability.

Again, Belgium is at odds with the European supranational and international positive obligations that require an effective prosecution of incitement to violence, discrimination, hate and segregation and disseminating ideas of racial superiority. The threshold of 'particular malicious intent' further adds a threshold to prosecution. But overall, apart from the legal requirements, there appears to be a lack of either knowledge on the framework of unlawful hate speech or lack of willingness to prioritise these cases. Further research needs to clarify the latter to better understand the low number of prosecution (or alternative actions). An important evolution to be followed in the future are the actions taken by several offices of the prosecution to develop a trajectory with external partners on hate crimes, including hate speech as an alternative to prosecution before court, such as the project by the Antwerp public ministry with Dossin Kazerne.

All in all, the study in work package 2 on OHS concludes that notwithstanding an all-encompassing legal framework on the criminalisation of incitement to discrimination, violence, hate and segregation, as well as the dissemination of racial superiority and hatred, filing a complaint for such offences is in practice often futile.

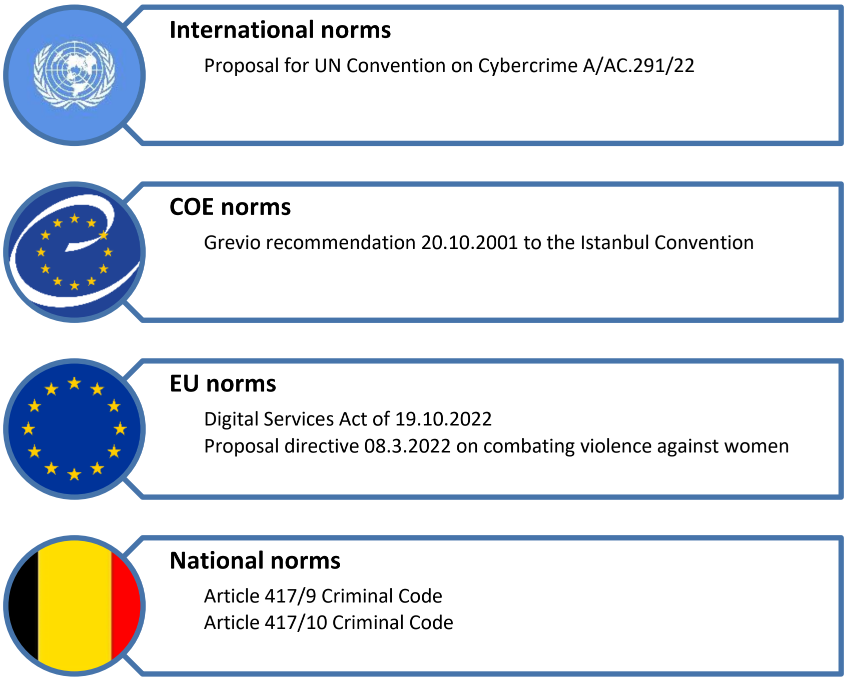
b. The legal framework of NCII

i. Mapping the legal framework of NCII

In line with the mapping exercise for OHS, the team mapped international, European supranational and national norms relevant to NCII on 3 levels, namely norms that hold explicit reference to NCII, norms that do not provide an explicit reference but are relied upon to tackle NCII, and finally, those rights and principles that are foundational for the norms on NCII (for the full mapping, see annex 7).

- **Level 1: international, supranational and national norms explicitly addressing NCII**

In the first place, those norms were mapped that explicitly concern the non-consensual dissemination of intimate images at the international, supranational and national level.



International norms

Proposal for UN Convention on Cybercrime A/AC.291/22

COE norms

Grevio recommendation 20.10.2001 to the Istanbul Convention

EU norms

Digital Services Act of 19.10.2022
Proposal directive 08.3.2022 on combating violence against women

National norms

Article 417/9 Criminal Code
Article 417/10 Criminal Code

Figure 6. Norms on NCII for mapping the legal framework

There, norms were mapped according to similar criteria as those mapping OHS.

Level of legislating	National, supranational (EU or Council of Europe), international
Binding	Binding norms / proposal for binding norms or non-binding (soft) law e.g., guidelines, …
Actor	Rules regarding NCII binding on state only (state obligations), natural and legal persons (e.g., criminalising certain behaviour) or other specific other actors (e.g., industry such as social media)
Criterion	Are the rules drafted in a general manner or do they focus on a specific personal characteristic of the victim, e.g., gender, age or sexual orientation?
Technology	Focus on or special consideration for online NCII (tech-specific) or on all forms of NCII (neutral)
Focus	Does the norm only include specific material, such as pictures or videos, or does it also address other forms of content, including manipulated content (e.g., deepnude)?
Intent	Is intent or harm a condition for the behaviour to be considered NCII, or is the lack of consent sufficient?

Table III. Categorisation criteria for NCII norms

The analysis shows that there are far less norms explicitly relevant to NCII in comparison to OHS. This is no surprise, given that NCII has sharply risen and become problematised due to digitalisation. Whereas several national states, including Belgium, already criminalised the dissemination of NCII, at the supranational and international level there are not yet binding rules explicitly concerning NCII, except for the DSA specifically addressed to online intermediaries. However, both at the international UN level and at the level of the EU, new norms are proposed that would address NCII.

When looking at the current and proposed norms, it is clear that most norms try to cover NCII in its broadest sense by taking a broad focus as to what material or content is targeted by the norm, including manipulated material, as well as opting for a technology neutral approach. The draft UN Convention on Cybercrime takes a tech specific approach, which follows from the focus of the Convention, namely on cybercrime instead of on gendered violence. The norms in general do not limit the scope of the norm to a specific group based on personal characteristics such as gender or age. Nevertheless, certain norms are clearly developed from a gender-based approach, e.g., whereby the preamble clearly refers to the protection of women and/or girls. However, this does not translate into a gender-specific norm, which means that these norms are applicable to the sharing of intimate images of women, men or other genders. The norms do not include a requirement of a specific intent or purpose on behalf of the perpetrator for criminalisation. Consent is included in these norms as the cornerstone to evaluate the behaviour as criminal or to be criminalised.

The above shows that there is as yet a limited set of binding rules on NCII. However, in so far the victim is a minor, this behaviour is also covered by international, European supranational and national rules on child sexual abuse material (CSAM). At the international, supranational and European as well as national level there is a vast developed framework on the dissemination of CSAM. In this respect, the dissemination of NCII of minors can be prosecuted on the basis of CSAM regulation and national states are under positive obligations to criminalise such dissemination. The difference between both frameworks of NCII and CSAM is the defining reason for the prohibition of the dissemination of the images, i.e., the lack of consent with regard to NCII and age with regard to CSAM (Gangi et al. 2022). Yet, minors are in principle considered not to be capable of consenting to the dissemination of intimate images, in consequence of which there will be an overlap with CSAM and NCII norms.

The national legislator exempted from CSAM the exchange of intimate images between consenting minors above 16, i.e., so-called sexting exemption (article 417/49 Criminal Code). However, in so far there is a lack of consent for the dissemination, there remains an overlap in national law between the provision on NCII and the provision on CSAM. Currently, there is no guidance for prosecution which norm to choose in such cases (Van de Heyning et al, 2023). However, this might have a huge impact as the social stigma is particularly attached to CSAM offending (Kothari, 2021). Moreover, the sentences differ in cases of CSAM or NCII, leading to an overlap in social stigmas and consequences. Therefore, it is important to have guidance in case of NCII of images of a minor.

- **Level 2: added layer of norms that do not explicitly mention NCII**

Secondly, the mapping exercise focused on those legal norms that do not explicitly mention NCII but are considered to protect against NCII. Like for OHS, literature and the court cases within the coding study were analysed to further map the second layer of the legal framework on NCII. Two such additional layers or norms protecting NCII were unearthed in the research.

First, courts, public organisations or scholarly research may read positive obligations on the state to protect victims of NCII in norms concerning the protection of privacy, personal and mental integrity, sexual rights, or equality and non-discrimination. The mapping study found several binding norms in which positive obligations to tackle NCII were read.

- The clearest example is the European Court of Human Rights that reads a positive obligation on states to protect individuals or a community against NCII based on the protection of the right to privacy and personal integrity entrenched in article 8 ECHR ((ECtHR 14 September 2021, *Volodina v. Russia n°2*). Moreover, the Court found that in combination with other physical forms of intimate partner abuse, NCII may amount to inhuman and degrading treatment, violating article 3 ECHR (ECtHR 9 July 2019, *Volodina v. Rusland n°1*).
- Positive obligations to tackle NCII are also read in obligations to protect women and girls against gendered violence. The expert committee to the Council of Europe's Istanbul Convention on gendered violence and intimate partner violence argued in its first additional recommendation that states are under a positive obligation to take measures on violence against women in the digital sphere. More implicitly, the CEDAW Committee recommended states in General Comment n°35 on the elimination of violence against women to take account and act against new forms of violence against women in the digital environments.

Second, national courts confronted with OHS can also rely on other norms that were not developed or intended to tackle NCII.

- Several criminal provisions can be applied to specific forms of NCII, e.g., article 146 §3bis of the Act on Electronic Communication on cyberharassment can be relied upon in case images are disseminated in order to cause harm, article 417/9 criminal code on voyeurism can be relied upon if the disseminated images were taken without the knowledge of the victim or digitally manipulated (deepnude), article 468 criminal code on extortion can be relied upon in case of sextortion.
- NCII can also often be tackled by personality rights, i.e., the protection of the right to image. Article XI.174 of the Belgian Code of Economic Law provides that a person's permission is requested to capture, exhibit, or reproduce his or her image.
- At the European supranational level, the right to image is included in a broader protection of the right to privacy and personal data. The protection of personal data resulted in a broad protection based on the ECHR (article 8 ECHR) and EU law (article 7 CFREU) and developed in EU secondary law, most notably the GDPR. The GDPR provides that factors specific to the physical identity of a person are to be considered personal data and can, therefore, only be processed with consent. This regulation also provides for the right to be forgotten, i.e., the right to have personal data erased, on the basis of which a victim could ask to have his or her image removed and/or the personal data that were posted along with it (e.g., in case of doxing, the making of material containing the personal data of another person without that person's consent, accessible to a multitude of end-users).

- **Level 3: fundamental rights restricting action against NCII**

The several documents explicitly or implicitly relevant to tackling NCII were analysed as to potential limitations regarding action against NCII. Different from OHS, there is in general no mention of fundamental rights or principles limiting the criminalisation of NCII. The only explicit mention as to limitations of NCII is in the Belgian 2020 law on the prohibition of NCII (Law of 4 May 2020 on the non-consensual dissemination of sexual images and recordings).

Here the preamble provides that in view of the rising number of NCII, a good balance should be worked out between freedom of expression and the right to protection of the privacy under the proportionality principle (Parl. proceedings Chamber 55-101). The preamble continues with mentioning the high percentage of women victimised by gendered cyberviolence, suggesting this to be a sufficient justification for NCII legislation. Also, with regard to the removal of images in view of the proposal of EU directive on gender-based violence, there is a reference to the freedom of expression, legality and proportionality in the explanatory Memorandum. Equally, in the preamble of the DSA (EU Digital Services Act) a general reference to the freedom of expression, right to information and freedom of press is made in view of removal and content moderation, albeit not specific to image-based sexual abuse such as NCII (Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC).

The explicit incrimination of NCII with reference to the freedom of expression in the Belgian legislation implies that the legislator found such incrimination a proportionate limitation of the freedom of expression and information. However, in most other documents there is no reference to the freedom of expression or other fundamental rights potentially limiting the criminalisation of NCII or obligation for removal of such norms. The hypothesis for this absence is that in general the freedom of expression, right to information and freedom of press is considered as not applicable to NCII. These rights include the protection of both text and images, but are only applicable in so far an opinion or information is disseminated. It poses the question whether this is the case with the dissemination of NCII. In conclusion, whether or not relevant, from the mapping of national, supranational European and international norms it is clear that the fundamental rights framework does not constitute a limitation for criminalising NCII.

- **Level 4: fundamental principles for shaping the criminalisation of NCII**

The mapping exercise of fundamental rights and principles showed a high level of coherence in the rights and principles invoked in the international, supranational and national documents relevant for the protection against NCII. It is clear that three sets of fundamental rights are central to shaping the protection against NCII: the protection of personal integrity (including physical, psychological and mental integrity), the protection of privacy and personal data, and the protection of equality and non-discrimination. The latter right goes along with the finding that, while most of the norms are drafted in a general manner and, therefore, applicable to all victims of NCII, these norms refer specifically to the gendered nature of NCII. Moreover, several of those norms in their title or preamble clearly indicate that the protection of NCII or more general digital gender-violence is approached from the perspective of the protection of women and girls.

In addition, several documents refer to the rights of the child and take into account age (minors or children) as an important element. The protection of these rights as framing the prohibition of NCII is shaped by the defining principles found in these documents, i.e., the protection against gender-based violence and equality. While consent features prominently as the crucial criterion in norms prohibiting NCII, it is remarkable that the protection of sexual integrity is only explicitly mentioned in the national incrimination of NCII. Sexual integrity is also mentioned in the proposal for an EU directive on gender-based violence, but only with regard to rape, suggesting that this principle would only be relevant for physical sex crimes.

ii. Coding of criminal cases on NCII

As the mapping exercise showed, the criminalisation of the non-consensual dissemination of intimate images is rather recent. In the case law study, the team got access to the cases at four prosecution offices (East Flanders, Halle-Vilvoorde, Brussels, and Namur) in the period 2018 - 2021 and at three courts (Liège, Brussels, and Antwerp) in the period 2016-2021. Unlike with discrimination and hate speech, there is not one code the prosecution internally uses to filter NCII cases. Therefore, the team filtered the cases on the basis of the several criminal offences known to be relied upon in prosecution for incidences of NCII, i.e., voyeurism, CSAM and harassment. Out of this total of select cases, 423 were selected as cases concerning NCII. The filtering exercise made it clear that there is a lack of clear categorisation within the database of the public ministry and the court system which makes it difficult to filter and analyse such cases. A specific categorisation for NCII, even though a separate criminal offense, is lacking. This means that not only for researchers, but also for the public prosecution and judges it is difficult to find and analyse such cases.

- **Number of complaints**

Given that the provision on NCII is only recent, it is clear that this is a burgeoning field of prosecution with a significant stock of cases. Moreover, there is an increase in cases over the period under examination. The research also shows a growing number of cases over the period under scrutiny. All but two cases were started by a complaint at the level of the police.

- **Reason for filing a complaint**

A remarkable finding is that most victims indicate that the main reason for filing the complaint is to have the intimate images deleted and to stop the suspect from disseminating the pictures. The second most common reason is to see damages and compensation. It appears from the cases that retaliation, i.e., punishment of the suspect for his or her actions, is of lesser concern for victims. This also shows how NCII requires a different perspective of law enforcement. It is within the DNA of law enforcement to focus on prosecution, i.e., finding evidence against a suspect to prosecute this person before a court or decide on an alternative measure. However, victims appear to file a complaint in the first place to seek help for preventing or stopping the dissemination of intimate images. This reason for filing a complaint is highly prevalent among victims of sextortion.

- **Discontinuation of prosecution**

This high level of incoming cases has, however, not (yet) translated to many cases before the criminal courts. The research found a high number of dismissals, with only 19 cases reaching the criminal court and just 15 cases terminated by the public office via other means such as mediation or probation. In the stock, the team found several comparator cases, i.e., cases with quasi-identical or comparable relevant facts, whereby in some cases there was further investigation, while in others this was not the case, or in some cases further prosecution, while in other these cases were discontinued.

Within the stock of those cases a wide variety of reasons can be found, including the lack of capacity for investigation, that it is not a priority for the internal case management of the prosecution, a disproportionality between the facts and prosecution, that the offenses are of relational nature, and a lack of evidence. When zooming in to the high number of cases dismissed for a lack of evidence, in many cases no further action is taken when the suspect refutes having disseminated the videos. As such, it appears that little further digital investigation is carried out in support of the victim's accusation.

The second most important reason for dismissal is that the suspect is unknown. In those cases as well, closer scrutiny shows that often no further investigation was carried out. It appears therefore that within the vast share of complaints dismissed for insufficient evidence or because the perpetrator was unknown, a lack of priority or capacity impeded further investigation and potential finding of evidence and/or a suspect.

- **Incriminations for prosecution**

When looking at the incriminations based on which the complaints were filed, the case study found a wide variety of incriminations, including hacking, harassment, CSAM, voyeurism, NCII, extortion, and more. Often, similar facts are being labelled under different incriminations. Moreover, it appears that in each case where an image of a minor is disseminated, the case is categorised as CSAM, even though the suspect is equally a minor and this is disseminated in a relational context. As such, it appears that there is little understanding of the full framework of norms and streamlining of the incriminations.

- **Social media used for NCII**

The vast majority of NCII is disseminated via Facebook (66), Instagram (65) and Snapchat (79). When taking a group-specific approach, the Meta group applications (with Messenger in 54 and Whatsapp in 28 cases) are most referenced to in the cases in the sample. A reason for this may be that these are simply the most used social media and communication applications in Belgium. This does not necessarily mean that most of NCII occurs on these social media, because victims might not be aware of the dissemination on more fringe or encrypted apps, e.g., on Telegram many groups exist for the dissemination of NCII. However, because these groups are member only and messaging is encrypted, victims might not be aware of it or believe filing a complaint is futile, as Telegram does not cooperate with law enforcement on this issue. These hypotheses will need to be further tested in future research.

- **Relation between the victim and perpetrator**

When looking at the connection between the victim and perpetrator, it is clear that technology-facilitated intimate partner abuse (TFIPV or tech abuse), i.e. abuse of digital means for intimate partner violence, is a common reason for NCII. In cases with a known perpetrator, 184 victims accused their (ex-)partner of having disseminated or threatened to disseminate the intimate images. The second most common connection is a virtual friendship, either amical, sexual, or romantic. Far less common are those cases where the images were disseminated by colleagues or school mates.

- **Reason for NCII**

The analysis of the complaints further showed remarkable findings as to the reason why the images were disseminated. The intent was either reported by the victim or by the suspect. This element has an important legal impact, as Belgian law holds a provision on NCII and one on NCII with malicious intent or for financial profit. The latter form of NCII is sanctioned with longer prison sentences and higher fines. The research found that in 124 cases, malicious intent was present. Even more common is profit, as the research found that in 177 cases, either the victim or the suspect mentioned that there was a financial input for NCII, i.e., either to make money by extorting the victim or to make money by disseminating the picture. As such, it should not come as a surprise that an important part of the cases concern sextortion cases, whereby the victim is threatened to pay a certain amount in order for pictures not to be disseminated.

The prevalence study among adolescents and emerging adults by no means found a similar high incidence of sextortion, nor did it find such a high level among respondents who have disseminated pictures for financial intent. Therefore, the hypothesis is that people will show more readiness to file a complaint with the police in case of sextortion or when they have lost money in order to prevent the extortion. As such, financial loss might be an important indicator for filing a complaint at the police. In addition to malicious intent (e.g., revenge) or financial motivation, perpetrators are also reported to have disseminated or shown intimate images to others for fun, sexual excitement, to prove sexual prowess, or to get other pictures in return.

- **Perpetrators and victims of NCII**

The vast majority of the suspects were male (287), whereas 68 suspects were female. In most cases with an identified suspect only one person was withheld, while in 20 cases there were multiple suspects. Most of the suspects were below 30 years old, with 96 of the suspects being minors, all but one suspect being teens and 145 of the suspects being below 30. The youngest suspect was 9 years old. Further, 94 suspects were in the age category 30 – 40, 37 in the age category 40 – 50 and 31 suspects beyond 50, with the oldest suspect being 68. As such, it is clear that age and gender are significant criteria for prevalence as a suspect in criminal complaints.

Whereas suspects are disproportionately more likely to be male, victims are disproportionately likely to be female. Within the case sample, 309 of the victims were female, whereas 114 victims were male. The latter number is, however, significant and works against the common perception that NCII victims are women only. In the sample of cases, the team found only a very limited number of multi-victim cases. This might be explained by the fact that almost all cases were started with a complaint filed by a victim. Victims will only focus in their complaint on their own context and facts or will be ignorant of other victims. In consequence and in view of the limited capacity of law enforcement, the investigation mostly focuses on the case of the victim and does not expand the case to other potential victims.

As to age, it is clear that there is a particularly high number of younger victims in the case sample: 133 victims were under 20, of which 99 victims were minors, and 116 victims were between 20 – 30. Far less complaints were filed by victims over 30 years old: 54 victims were aged between 30 – 40 and 60 beyond 40. This conforms with other research on the topic that shows that age is the most important predictor for victimisation of NCII.

iii. Analysis of case law research in the light of the mapping exercise

The coding exercise shows that there is an increasing number of complaints with regard to NCII at the level of law enforcement. However, only a limited part results in actual prosecution. When zooming in, the reason is that often, there is no or limited investigation into these cases. The research hypothesises that this might be due either to a lack of capacity and prioritising and/or a lack of knowledge as to the investigation into this crime of online image-based sexual abuse. This is further supported by the wide variety of incriminations relied upon in cases of NCII. This is remarkable because there is a clear incrimination in national law on NCII. Again, this shows that there is a lack of knowledge and guidelines in this perspective, particularly apparent concerning NCII among minors, where both incriminations of NCII and CSAM are used incoherently.

The coding exercise further supports that gender and age are indicators for a higher prevalence of complaints in NCII. Remarkable from the coding is that forms of sextortion and NCII within an intimate partner context are highly prevalent. It suggests that victims are more forthcoming in filing a complaint when a financial loss, a financial threat, or an element of intimate violence is present. Moreover, the main reason to file a complaint is to stop or prevent the dissemination of the images or the accompanied harassment. The punishment of the perpetrator appears to be only a secondary motive.

The case study shows that in general, when comparing with the results from WP3 on prevalence, only a small number of victims file a complaint. Several hypotheses might explain this gap. A first hypothesis is a lack of knowledge on the criminal nature of NCII. However, this hypothesis can generally be disregarded, as WP3 shows that the vast majority of victims thinks this behaviour should be criminally sanctioned.

A second hypothesis is the lack of belief in the capacity of law enforcement in investigating and prosecuting such crimes, which was also mentioned in the analysis of the coding on OHS (Van de Weijer, 2020). The high level of discontinuation of complaints would further support such a hypothesis, especially given that in many of the cases there is no further investigation. As such, in order to have more prosecutable cases, it appears important to communicate better regarding which evidence should be included when filing a complaint. Moreover, more knowledge, capacity, and training for law enforcement on investigating such cases might be beneficial. This has become all the more urgent during the research, as it was communicated that the specialised unit of the federal police (IRU-I2 of the DJSOC) handling cases of NCII and in particular their removal, will be included in the general and already understaffed Federal Computer Crime Unit and will no longer provide assistance to the local police. This will further impact the capacity and skills for addressing these cases.

A third hypothesis might be that the main reason for action is the removal or prevention of the dissemination of these images or halting the accompanied harassment of extortion. It could be theorised that, as the objective of a criminal complaint is generally identifying, prosecuting and punishing the prosecutor, this might not be the first objective of victims and that, therefore, they do not file a complaint to the police but first consider other options. Further research into the reasons for filing a complaint can clarify the thresholds for filing a complaint.

c. OHS and NCII in the law and the courts: comparative findings

The results from the mapping exercise on OHS show that there is an all-encompassing framework at the international, European supranational and international level to delineate which forms of OHS are to be considered lawful or unlawful hate speech. This framework exists of international and supranational norms that hold positive obligations for states to prevent, prohibit or criminalise hate speech on specific grounds, particularly on the grounds of race, skin colour, ethnicity, nationality, religion, and gender. These positive obligations translate at the Belgian national level to the criminalisation of hate speech on all grounds relevant to a person’s status or characteristics, albeit based on different legislative norms and further supported by generic provisions of the criminal code, e.g., provisions on cyber harassment. The vast majority of these rules are technology neutral. However, in recent years, more specific references to the impact of digitalisation or even cyberhate provisions are emerging.

When drafting the legal framework on the delineation of lawful and unlawful OHS, this might diverge based on the specific characteristic or status targeted by the specific expression of hate, e.g., race, gender, sexual orientation, or disability. Research finds that the most all-encompassing framework on hate speech exists with regard to racism and xenophobia (race, skin colour, ethnicity, nationality) and religion, and to a lesser extent to gender. The legal framework is less developed for hate speech based on other personal characteristics and status. Moreover, these norms do not focus on the same actors. They may be construed as positive obligations on states, criminal provisions on natural and legal persons or obligations on the digital industry (online platforms, intermediaries, IT companies).

The delineation of what speech can be criminalised is at the national, European supranational and international level formed by the freedom of expression, either by an explicit reference within the document or norm on OHS or by human rights treaties and constitutional norms. When looking at the interpretation of these norms in the context of hate speech, there are several cases where the importance of the ‘internet’ or ‘digital’ component is highlighted for drawing the line between lawful and unlawful speech. For this delineation, a further layer is added by values and principles in the light of which the positive obligations to prohibit and sanction hate speech are balanced with the freedom of expression and right to information. Such values are particularly prominent in the documents at the European supranational and international level, where it is clear that the key values consist of freedom, equality, democracy, pluralism, tolerance and respect. Security is a more common value when considering hate speech based on gender or sex.

In conclusion, the delineation of lawful and unlawful speech is a delicate balancing of several rights and interests, where judges need to consider a multi-layered legal framework specific to the relevant grounds of hate speech in the case at hand as well as to the actors concerned (states, natural & legal persons, industry).

In contrast to OHS, there is still a limited setting of binding norms at the international and European supranational level on the prohibition of NCII. This criminalisation is also important for the removal of such images, as this would equate NCII as unlawful content to be removed by digital intermediaries under the DSA or other norms relevant to content moderation and removal. However, while such binding norms are still lacking at the international and European supranational level, it is clear from the mapping exercise that other norms might also cover NCII to define NCII as illegal speech, most notably norms on CSAM in so far the victim is a minor, the protection of privacy and personal data, the protection of the right to image, the protection of personal integrity and principle of gender-based violence as a principle of international customary law and the protection of equality and non-discrimination. As such, there appears already a clear framework of rights and principles shaping a sound basis for an international and supranational protection against NCII.

The concrete binding norms on criminalisation are as yet to be found at the national level. Belgium, like several other countries, has criminalised NCII. As such, it appears that the driving force for criminalising NCII, contrary to OHS, is the national level, while the international and European supranational level is now following up on this evolution with the draft convention on cybercrime at the level of the UN (Proposal for UN Convention on Cybercrime A/AC.291/22) and proposal of directive on gender-based violence at the level of the EU (Proposal for a directive of the European parliament and the council on combating violence against women and domestic violence COM(2022) 105 final 2022/0066 (COD)). As such, an international and European framework is under construction.

Two issues arise from the mapping exercise to be addressed in the future. First, there is the clear overlap with CSAM norms. Belgium has already taken a step forward with the sexting exemption for minors sexting with consent, but there is still a potential for overlap on all levels of regulation. Better defining and delineating NCII from CSAM would be beneficial for coherence, prevention, and treatment. Second, the fundamental rights framework on delineating the limits of NCII is underdeveloped. References to the freedom of expression and proportionality are only found in a limited number of documents and not fully addressed. However, whereas on the one hand one can question whether NCII in general falls within the scope of this right, drawing the line between art and satire on the one hand and digitally manipulated images denuding victims might warrant such discussion.

When looking into the coding of the case law for OHS and NCII, there are quite some parallels. First, there is a low tendency to file complaints in comparison to the prevalence of these behaviours online. Second, in the vast majority of those complaints, there is either no further investigation or the case is discontinued, even in cases with a known perpetrator. Only a handful of cases will be prosecuted before the courts. The research found several comparator cases in the stock of cases of OHS and NCII where in some cases, there was investigation and prosecution while in the quasi-identical cases there was no investigation and prosecution.

The team developed several hypotheses for this low number of complaints and high level of discontinuation. A common hypothesis for both OHS and NCII is the potential low trust of victims in law enforcement to be able to understand and investigate online phenomena prevalent on social media, in particular new trends and behaviours. This lack of trust can be further amplified by the high number of discontinuations. From the case study it appears that lack of capacity or priority for these cases play an important role in the discontinuation.

A hypothesis specific to OHS is that the procedural threshold in Belgian criminal law for prosecuting hate speech, i.e. the competence by the Court of Assize for all cases of public OHS with the exception of racism and xenophobia, results in a pre-filtering of complaints, in consequence of which complaints will mostly be filed for OHS on the basis of gender, race or nationality.

A hypothesis specific to NCII is that victims might not consider filing a complaint as their main concern is the prevention, removal, or discontinuation of the dissemination of intimate images rather than the investigation, prosecution, and punishment of the perpetrator. As law enforcement is generally more associated with the latter, victims might be less inclined to file a complaint for NCII, even though law enforcement is particularly tasked under Belgian law to help victims in removing such images.

The team also found a wide variety in incriminations used to cover OHS and NCII that are not always the best fit. Further knowledge and training of law enforcement on this point appears important. But victims (or bystanders) are also not sufficiently knowledgeable of the distinction between lawful and unlawful conduct regarding OHS and NCII, thereby filing complaints for behaviour of OHS or alleged NCII that is not punishable under the Belgian law. The fact that there are only a few court decisions on OHS and NCII, which are scarcely published and difficult to find in the databases, amplifies the lack of knowledge and foreseeability.

3.3 SURVEY AND VIGNETTE STUDY ON ONLINE HATE AND NCII AMONG DIGITAL NATIVES

3.3.1. METHODOLOGY

a. Participants & Procedure

In collaboration with a research agency (Profacts), an online survey (in Dutch and French) was conducted among late adolescents and emerging adults (15 to 25 years old) (see annex 8). To guarantee a representative sample in terms of language and gender, we based ourselves on the Belgian prevalence rates of Statbel:

- *Quota gender*
  50,9% (663 293/ 1 301 827) of the Belgian adolescents and emerging adults (15-25 years old) were male in 2021. To meet these rates, our survey must count 1434 males (Statbel, 2021). 49,04% (638 534/1 301 827) of the Belgian adolescents and emerging adults (15-25 years old) were female in 2021. To meet these rates, our survey must count 1382 females (Statbel, 2021).

- *Quota Dutch-speaking/French-speaking*
  50% of the respondents speaking Dutch (1409) and 50% of the respondents speaking French (1409).

- *Quota LGBTQIA+ community and respondents with a foreign background (Belgian and non-Belgian)*
  As the panels of the recruitment agency have a low number of respondents with a foreign background, they executed a boost of n=200 to guarantee enough respondents with a foreign background. The overall prevalence rates of the LGBTQIA+ community are 3%, estimated by the recruitment agency, which counts for 141 respondents.

The data was collected at two periods in time. The first wave was in January 2023 (N=1819) but as the research the team included an extensive experimental vignette study that needed a higher number of respondents, the research team decided to send the online survey to another pool of respondents in April 2023 (N=1000). In total, 2819 respondents participated in the survey (M*age* = 20.50 years, SD = 2.9).

As the study investigated sensitive topics among young respondents, the team informed potential respondents about the content and the aim of this study, and asked for explicit consent to participate. Respondents younger than 16 years old cannot directly be recruited; therefore, their parents were approached by the research agency to give their informed consent for their children to participate in this survey. Profacts works with their own reward system, where respondents can earn "points" which can be used later either to trade for a small amount of money or to donate to a charity. Due to the sensitivity of the topic, all respondents were provided with an information sheet that refers to formal support organisations at the end of the survey. The complete procedure was submitted and approved by the Ethics Committee of Social Sciences and Humanities (EASHW) of the University of Antwerp.

b. Measures

To analyse the data, and to measure the differences between Chi-square tests, all variables that measured socio demographics were transformed into categorical variables.

- Age was transformed in a two-category variable: adolescents (15-17 years old) and emerging adolescents (> or = 18 years old) as based on other research taking age into consideration (Gassó et al.; Pedersen et al., 2022; Van Ouytsel et al., 2017).

- Ethnicity was transformed in a two-category variable: Belgians (respondent, mother and father are born in Belgium), Belgians with a foreign background, non-Belgians with a foreign background. This division follows the rules of the Statbel data (Statbel, 2023).
- Gender was transformed in a three-category variable: men, women and other (i.e., non-binary and transgender). The category "other" was not further split up to make sure the sample size would be big enough for the further analysis.
- Sexual orientation was transformed into a two-category variable: heterosexuals and LGBTQIA+. LGBTQIA+ was not further split up to make sure the sample size would be big enough for the further analysis (Meechan-Rogers et al., 2021).

To map the sociodemographics of our sample, the survey included questions regarding age, gender, sexual orientation, and ethnicity (see annex 8, Q2-Q8). The possible answers were equal to those used in WP1 to guarantee continuity. The final sample was composed as follows (a graphic presentation of the sample composition is presented in annex 9):

- 46.5% of the respondents are men (N=1312), 51.7% are women (N=1457) and 1.8% belongs to the subgroup "other" (i.e., transgender people and non-binaries) (N=50).
- 83.4% of the respondents self-identified as heterosexuals (N=2350) and 13% as LGBTQIA+ community (N=379).
- 65.8% (N=1856) are Belgian with no mention of a foreign background, 23.9% of the respondents are Belgians with a foreign background (N=673; mother or father born abroad) and 10.3% of the respondents are non-Belgians (N=290; respondent born abroad).

i. Scales for measuring prevalence rates

The research team collaborated with a research agency that has contact with several respondent panels. Due to the sensitivity of the topic, there was one panel that refused to include questions that map potentially illegal behaviour (marked as "hide if source is 3" in annex 8). This panel included 551 respondents and therefore the prevalence rates (victimisation, perpetration and bystandership) of online hate speech and NCII were only measured in a sample of 2268 (2819-551=2268) respondents.

Online hate speech

Online hate speech is based on specific characteristics of the individual, namely gender, sexual orientation, and ethnicity. Therefore, it is important to measure the prevalence of these three types of online hate speech. There exists no validated scale to measure online hate speech. Therefore, the research team based the survey questions measuring online hate speech on previously conducted research to construct a scale measuring different types of online hate speech (Bedrosova et al., 2022; Hawdon et al., 2015). Two new variables were created out of these three types of online hate speech to have an overall idea whether online hate speech was prevalent or not:

- Total online hate speech victimisation is a two-category variable (yes/no) that indicates whether the respondent has been victim of at least one of the three subtypes of online hate speech (i.e., based on gender, ethnicity and sexual orientation).
- Total online hate speech perpetration is a two-category variable (yes/no) that indicates whether the respondent has already posted or sent at least one of the three subtypes of online hate speech (i.e., based on gender, ethnicity, and sexual orientation).

The research team included several questions to measure the prevalence of victimisation, perpetration, and bystandership of online hate speech (table I, annex 10). Online hate speech can be directed towards one person or a specific group (e.g., a woman or towards women in general).

The non-consensual dissemination of intimate images (NCII)

The prevalence of victimisation and perpetration of the non-consensual dissemination of intimate images (NCII) was measured by applying the "IBSA Perpetration Scale", which includes several types of NCII (table II, annex 10). From the interviews in WP1 and previous research (Henry et al., 2021; McGlynn et al., 2020; Powell et al., 2019; Powell et al., 2020), it appears that the possibility to identify the victim on the naked picture also plays a role and, therefore, the team added the following answering options: (part of) the face is visible, personal features of the body are visible (e.g., tattoo, scars, birthmark), personal characteristics of the environment are visible (e.g., the bedroom). Two new variables were created out of these scale to have an overall idea of NCII victimisation and perpetration:

- NCII victimisation is a two-category variable (yes/no) that indicates whether the respondent has been victim of at least one type of NCII (independently of the content of the image).
- NCII perpetration is a two-category variable (yes/no) that indicates whether the respondent disseminated an intimate image (independently of the content of the image).

To map bystandership of the non-consensual dissemination of intimate images, the research team extracted these rates from a follow-up question on the non-consensual dissemination of intimate images. In table III (annex 10), answering option four and five give a clear view on how many people ever received a naked picture of someone else.

Descriptive data and chi-square tests were used to analyse the prevalence rates of these two behaviours and to measure the difference within each diversity variable regarding victimisation and perpetration of both behaviours (i.e., gender, sexual orientation, ethnicity, and age). The results on all chi-square tests are presented in annex 11.

ii. Vignette Study

Content of the vignettes

For each behaviour, we developed eight vignettes to test a) the importance of diversity variables in victims, b) to measure the prototype willingness model, c) to investigate the harmfulness and d) to map the appropriate types of legal action according to the respondents. The vignettes on online hate speech are presented in annex 8 (p. 1-9 for the Dutch version and p. 60-67 for the French version). Scenarios A (SC1,SC2,G1,G2,G3,SO1,SO2,SO3) are focused on OHS and Scenarios B (SC1,SC2,G1,G2,G3,SO1,SO2) are focused on NCII.

For the content of the vignettes, the research drew upon the findings of the interviews focused on the qualitative understanding. Following the interviews, a table was created to systematically organise and analyse the collected data to have an overview to develop the scenarios. This table included a description of each case, including relevant details about the victim and perpetrator, such as their age, sexual orientation, gender, cultural background, and other pertinent characteristics. Information about the presence or absence of bystanders, their relationship with the other individuals involved, and their reactions to the behaviour were also recorded. The perceived or described intention of the perpetrator, as well as the contextual factors surrounding the incident (e.g., location, involvement of other people), were documented. Additionally, the impact of the behaviour on the victims and their coping mechanisms, as well as relevant psychosocial mechanisms, were noted in the table. This table was used for generating scenarios that will be utilised for the subsequent vignette study, allowing for a more controlled and comparative analysis of participants' responses.

In the vignettes, we focused on three diversity variables, namely gender, sexual orientation, and ethnicity. We manipulated each variable in the victim of each scenario to measure if this would result in differences regarding harmfulness and steps for legal action. Additionally, by investigating the profile of the respondent, i.e., their gender, sexual orientation, and ethnicity, potential differences in reactions concerning the constructs of the prototype willingness model (PWM) were investigated. In the following tables, the vignettes on respectively OHS and NCII are described including the sample size for each vignette. Wave 2 of the survey was organised to make sure that each vignette had enough respondents so that a) the analyses concerning the prototype willingness model could be done and b) to investigate the harmfulness of behaviour and the appropriate types of legal action.
The content of the scenarios was based on the experiences of the participants in WP1, the legal cases in WP2 and previous research.

Vignette Study to test the PWM
To test the prototype willingness model, the participants were presented with a series of questions that were previously used in research applying the PWM and followed the guidelines of Gibbons and Gerrard (Gibbons & Gerrard, 1995; Gool et al., 2015; Van Ouytsel et al., 2020; Walrave et al., 2015). Additionally, the team took into consideration the results of WP1 and even added items based on these results. Table V (annex 10) presents the items that were used to test the PWM's constructs.

As some questions try to map respondents' willingness and intentions to engage in potentially harmful and illegal behaviour, there was one panel that did not want to include the questions measuring respondents' willingness and intentions (N=551). To test the Prototype Willingness Model, structural equation modelling was applied. The modelling will present two models, one for online hate speech and one for NCII. In the model testing, the analysis will measure the influence of the three diversity variables on both NCII and OHS whilst also measuring the influence of previous perpetration (Q48, Q52,Q56, Q103 in annex 8) and financial stress (Q13-15 in annex 8).

Vignette study to investigate harmfulness and digital natives' perspective on legal action
To map the harmfulness of the two behaviours (online hate speech and NCII) and to measure the association of victims' characteristics in the vignettes and respondents' profiles, analyses were done on the item which measured respondents' attitudes (table V, see annex 10). The respondents had to answer this question by selecting an answer on a five-point Likert-scale. To map the harmfulness of a particular behaviour, the data were transformed by the answering options, where "Totally disagree" and "disagree" were interpreted as "not harmful", "nor agree, nor disagree" as having a neutral opinion on the vignette, and "totally agree" and "agree" were considered as stating that the depicted situation was considered "harmful".

The final follow-up questions of the vignettes were used to test the respondents' knowledge about the illegality of certain behaviour. Additionally, they were asked if they think this behaviour is illegal according to the Belgian penal code and which punishment would be best to put in place to sanction these behaviours. Table VI (annex 10) presents the follow-up questions used to map digital natives' knowledge on the illegality of the two behaviours and how they would structure the penal code. The interviews of WP1 gave the research team a first insight into how digital natives would legally approach both behaviours and were therefore used as inspiration for the answering options.
Because each vignette is followed by the same questions, it is possible to compare these features whilst considering victims' characteristics and respondents' characteristics.

3.3.2. RESULTS

a. Findings on prevalence

i. Prevalence of online hate speech victimisation and the role of diversity variables

The survey shows that a third of the digital natives already received online hate speech based on gender, sexual orientation, or ethnical background (34%), which is indicated in the following Figure 7 with 'total'. The category total indicates that the respondent received at least one type of online hate speech. More particularly, the survey revealed that 34% of the respondents (N=2268) had already received online hate speech, independently of whether this was based on gender, sexual orientation, or ethnicity. However, the prevalence rates also indicate that more digital natives received online hate speech based on ethnical background (24.2%) than online hate speech based on gender (14.6%) or sexual orientation (14.1%).

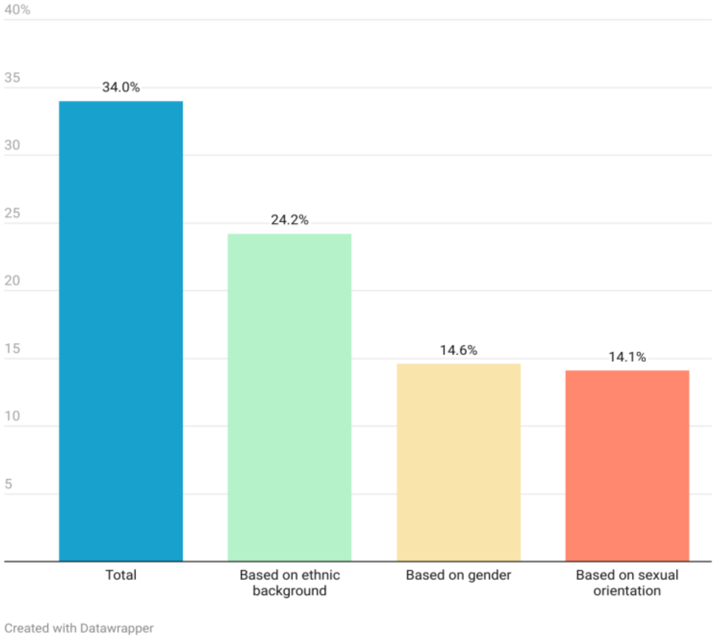


Figure 7. Prevalence of online hate speech victimisation

The aim of the survey was also to investigate if diversity variables (age, gender, sexual orientation, and ethnicity) are differing within online hate speech victimisation. The research team conducted Chi-square tests to investigate diversity variables and online hate speech victimisation and to measure any differences. A p-value ≤ 0.05 indicates that there are significant differences within the diversity variable and thus plays a role in online hate speech victimisation or perpetration. The relevant tables concerning the chi-square tests are presented in annex 11.

The division of the different subgroups for OHS within the diversity variables is the following: The sample (N=2268) exists of 46% males (N=1072), 52.5% females (N=1155), and 1.5% the subgroup "other" (N=41) (i.e., transgender people, non-binary). The sample (N=2200) consists of 86.1% heterosexuals (N=1894) and 13.9% members of the LGBTQIA+ community (N=306). The sample (N=2268) is composed of 65.7% respondents who are born in Belgium (N=1894; both the respondent and parents are born in Belgium), 24.1% respondents are Belgians with a foreign background (N=547; respondent is born in Belgium but mother and/or father are born abroad) and 10.2 % are non-Belgians with a foreign background origin (N=231; respondent is born abroad). The sample (N=2268) consists of 82.3% emerging adults (N=1867; > or =18 years old) and 17.7% adolescents (N=401; < 18 years old).

The role of gender

Table I (annex 11) presents the prevalence rates per type of online hate speech according to gender. Overall, the results show that men are more confronted with the three different types of online hate speech than women whilst "total" results show that women are more often than men victims of at least one type of online hate speech. One fifth in the subgroup "other" has ever received online hate speech based on gender. When looking at "total", one third of the respondents was victim of at least one type of online hate speech whilst this is 50% for the subgroup "other".

Chi-square tests showed that there are no significant differences (p-value > 0.05) in terms of gender when it comes to the different subtypes of online hate speech. However, there is a significant difference in the gender subgroups for "total" of online hate speech (p<0.05). Thus, respondents of the subgroup "other" receive more online hate speech (50%) than women (35.1%).

The role of sexual orientation

Table II presents the prevalence differences per type of online hate speech according to sexual orientation. Overall, the results show that both heterosexuals (23.8%) and members of the LGBTQIA+ community (28.8%) are mostly confronted with online hate speech based on ethnic background.

Chi-square tests showed that there are significant differences (p-value < 0.05) in terms of sexual orientation and gender when it comes to online hate speech based on gender and sexual orientation. Members of the LGBTQIA+ community receive more online hate speech based on gender (20.9%) and on sexual orientation (26.3%) than heterosexuals (13.5%, 12.2%).

The role of ethnicity

Table III shows that one out of five respondents of each ethnic subgroup is confronted with online hate speech based on ethnic background.

Chi-square tests showed that there are no significant differences (p-value > 0.05) in terms of ethnicity when it comes to the different subtypes of online hate speech. However, there is a significant difference within the subgroups of ethnicity for the group "total" of online hate speech (p<0.05). Belgians with a foreign background (41.3%) are most confronted with at least one type of online hate speech.

The role of age

Table IV presents the prevalence differences per type of online hate speech according to age.

One out of ten adolescents is confronted with all three different subtypes of online hate speech. When looking at online hate speech based on gender, the prevalence rates double when comparing adolescents (8%) to emerging adults (16%).

Chi-square tests showed that there are significant differences (p-value < or = 0.05) in terms of age in the three subtypes of online hate speech and the group "total". This means that emerging adults receive more online hate speech based on ethnic background, gender, and sexual orientation in comparison to adolescents.

Short summary on the role of diversity variables in online hate speech victimisation

In the abovementioned figures, some analyses show significant differences between diversity groups. For members of the LGBTQIA+ community, the prevalence rates of online hate speech based on gender and sexual orientation are higher. Women have received more of at least one type of online hate speech than men. Belgians with a foreign background also more often become victims of at least one type of online hate speech. In the three subtypes of online hate speech and the group "total", the prevalence rates are higher in emerging adults than in adolescents.

ii. Prevalence of online hate speech perpetration and the role of diversity variables

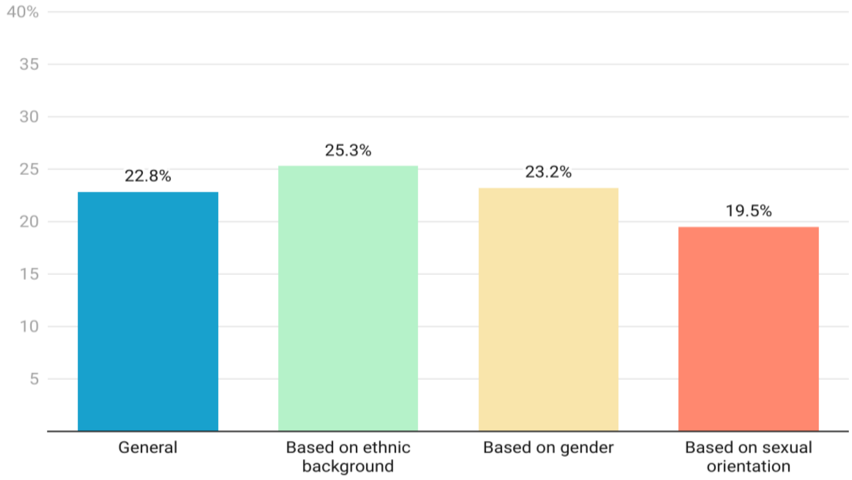
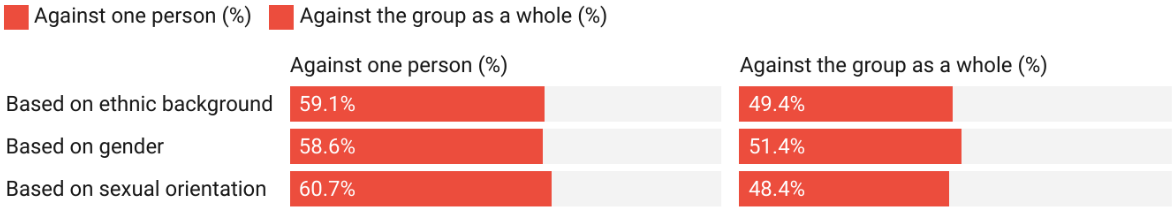


Figure 8. Prevalence of online hate speech perpetration

The survey shows the prevalence rates of digital natives who sent or posted online hate speech based on gender, sexual orientation, and ethnic background (see Figure 8). The survey revealed that 22.8 % of the respondents (N=2268) has already posted or sent online hate speech independently if this was based on gender, sexual orientation, or ethnicity. The results also reveal that digital natives posted or sent less online hate speech that is based on sexual orientation (19.5%). Figure 9 shows the percentage of offenders that directed their hate speech towards one person or the group as a whole (e.g., one woman or women in general). Within the context of all types of online hate speech, the results indicate that more offenders direct their hate speech against one person than towards a group in general.



Created with Datawrapper

Figure 9. Online hate speech against one person or against the group as a whole

The role of gender

Table V (annex 11) shows that one out of four of the men and women posted online hate speech, based on ethnic background and gender or, in general, at least one type of online hate speech. One out of ten respondents of the subgroup "other" posted all types of online hate speech. Chi-square tests showed that there are no significant differences (p-value > 0.05) in terms of gender in online hate speech perpetration.

The role of sexual orientation

Table VI shows that one out of four heterosexuals has sent online hate speech based on ethnic background. One fourth of heterosexuals has posted online hate speech based on gender, sexual orientation, or at least one of the three subtypes of online hate speech. One out of five members of the LGBTQIA+ community has ever posted online hate speech based on gender whilst one out of four posted online hate speech based on sexual orientation and ethnic background. Chi-square tests showed that there are no significant differences (p-value > 0.05) in terms of sexual orientation in online hate speech perpetration.

The role of ethnicity

Table VII shows that one out of five of each ethnicity group posted online hate speech based on ethnic background and gender. One of four of each ethnicity group posted hate speech on sexual orientation. Chi-square tests showed that there are significant differences (p-value < 0.05) in terms of ethnicity in the group "total". This means that non-Belgians with a foreign background posted at least one type of online hate speech the most (25.5%).

The role of age

Table VIII presents the prevalence differences per type of online hate speech according to age. One out of ten adolescents posted online hate speech based on sexual orientation or at least one of the different types of online hate speech. One out of four emerging adults posted online hate speech based on ethnic background, gender or at least one of the three types of online hate speech ("total").

Chi-square tests showed that there are significant differences (p ≤ 0.05) in terms of age when it comes to the different subtypes of online hate speech. In other words, more emerging adults post all the different types of hate speech than adolescents. Moreover, results show that hate speech based on ethnic background is sent the most, both by adolescents and emerging adults. When taking all types of online hate speech into account, prevalence rates are higher among emerging adults.

Short summary on the role of diversity variables in online hate speech perpetration

In the abovementioned results, some analyses show significant differences between diversity groups. Non-Belgians with a foreign background post the most online hate speech of at least one type of online hate speech. In the three subtypes of online hate speech and the group "total", the prevalence rates are higher in emerging adults than in adolescents.

iii. Prevalence of bystandership of online hate speech

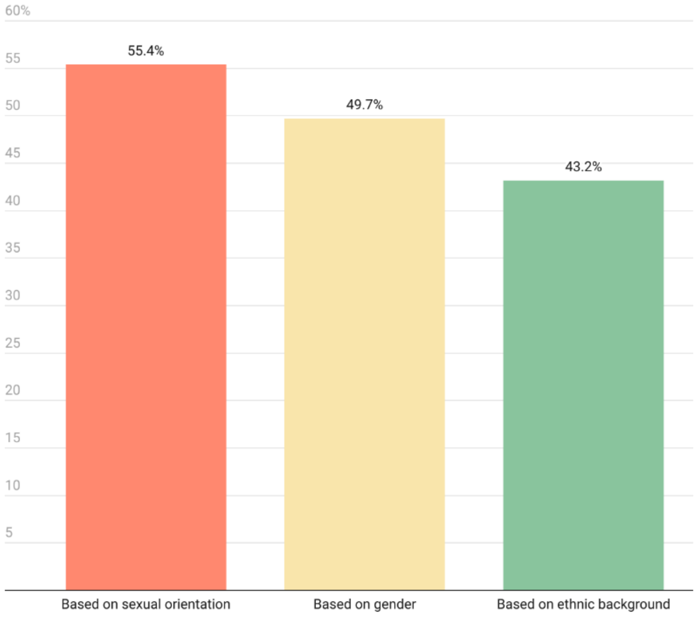


Figure 10. Prevalence of bystanders in online hate speech

The results show that more than 40% of the respondents (N=2668) have ever witnessed online hate speech. More than half of the respondents (55.4%) indicated to have ever seen online hate speech based on sexual orientation, while a little less than 50% has ever seen online hate speech based on gender (49.7%) and on ethnic background (43.2%).

iv. Prevalence of NCII victimisation and the role of diversity variables

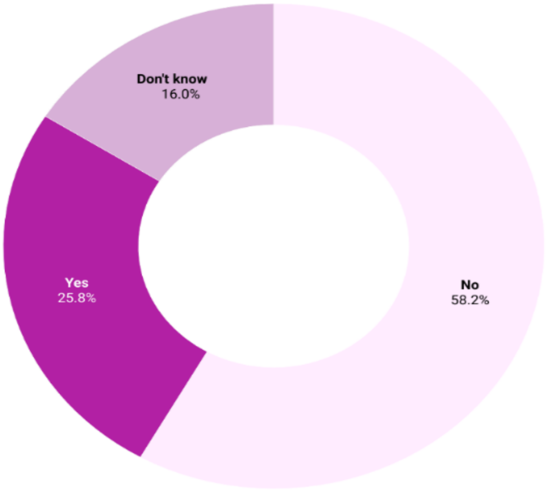


Figure 11. Prevalence of NCII victims

In total, 25.8% of the respondents (N=736) had their intimate image disseminated without their consent to others. 16% indicated that they don't know whether their intimate image has ever been spread or not. A majority of respondents (58.2%) state that their intimate image was not further disseminated.

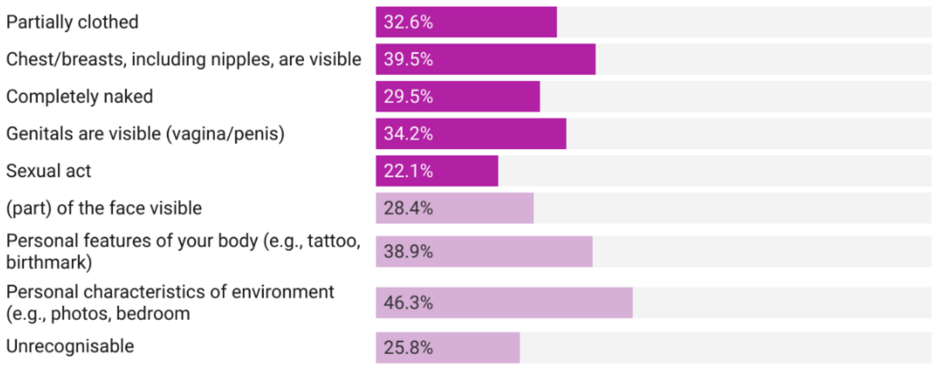


Figure 12. Content of the intimate image and victims' identifiers visible on the image

As far as the specific content of the intimate image that was spread is concerned, more or less one-third of the respondents were partially clothed (32.6%), showed their chest or breasts (39.5%) or showed their genitals (34.2%). In total, 29.5% of the respondents indicated that they were completely naked on the images that were non-consensually spread and in 22.1% of the cases, the intimate images captured a sexual act.

Almost half of the respondents (46.3%) indicated that personal characteristics of the environment were visible on the intimate image that was spread. More than one-third of the intimate images that have been spread captured personal features of the respondent's body (e.g., tattoo). 28.4% of the victims showed (a part of) their face on the intimate image that has been spread whilst in 1 out of 4 of the cases, the respondents were unrecognisable (25.8%). The aim of the study was to further investigate the role of gender, sexual orientation, ethnicity, and age in the context of NCII victimisation. The chi-square tests related to these are shown in table IX to XII (annex 11).

The sample sizes of the subgroups of all diversity variables within the context of NCII are the following: The sample (N=736) is composed of 46.3% men (N=341), 52.5% females (N=386) and 1.3% the subgroup "other" (N=9) (i.e., transgender people, non-binary). The sample (N=954) comprises 67.7% heterosexuals (N=646) and 32.3% members of the LGBTQIA+ community (N=308). The sample (N=736) involves 64.4% Belgians (N=474; both the respondent and its parents are born in Belgium), 24.6% Belgians with a foreign background (N=181; respondent is born in Belgium but mother and/or father are born abroad) and 11 % non-Belgians with a foreign background (N=81; respondent is born abroad). The sample (N=736) is composed of 90.6% emerging adults (N=667; > or = 18 years old) and 9.4% adolescents (N=69; <18 years old).

The role of gender

Table IX shows that in each gender group, one out of five has become victim of NCII. Chi-square tests showed that there are no significant differences (p-value > 0.05) in terms of gender in NCII victimisation.

The role of sexual orientation

Table X shows that both heterosexuals and members of the LGBTQIA+ community have been victimised.

Chi-square tests showed that there are no significant differences (p-value > 0.05) in terms of sexual orientation in NCII victimisation.

The role of ethnicity

Table XI shows that 27% of both the Belgians with a foreign background and non-Belgians with a foreign background have been victim of NCII. 24.9% of the Belgians has ever been victimised. Chi-square tests showed that there are no significant differences (p-value > 0.05) in terms of ethnicity in NCII victimisation.

The role of age

Table XII shows that 26% of emerging adults has even been victim of NCII, whilst 23.2% of the adolescents became victim of NCII. Chi-square tests show that this difference is statistically significant.

Short summary on the role of diversity variables in NCII victimisation

In the above figures, some analyses show significant differences between diversity groups. There are no significant differences in NCII victimisation in terms of gender, sexual orientation, and ethnicity. There is a significant difference between adolescents and emerging adults in NCII victimisation, meaning that more emerging adults have been victim than adolescents.

v. Prevalence rates of NCII perpetration and the role diversity variables

Figure 13 reveals that almost of 32% of all the respondents (N=688) disseminated an intimate image without the consent of the person depicted on it. One out of four NCII perpetrators indicated that they have disseminated pictures where the victim was completely naked (25.2%), chest or breasts were visible or where genitals are visible. One fifth of the perpetrators revealed that they have already disseminated intimate pictures of others without their consent where the victims were partially clothed (19.5%), a sexual act was performed (19.5%), (part of) the face was visible (21.9%) or where personal features (18.4%) or environmental characteristics are visible (19.1%). In one tenth of the disseminated intimate images the victim was not recognisable (9.3%).

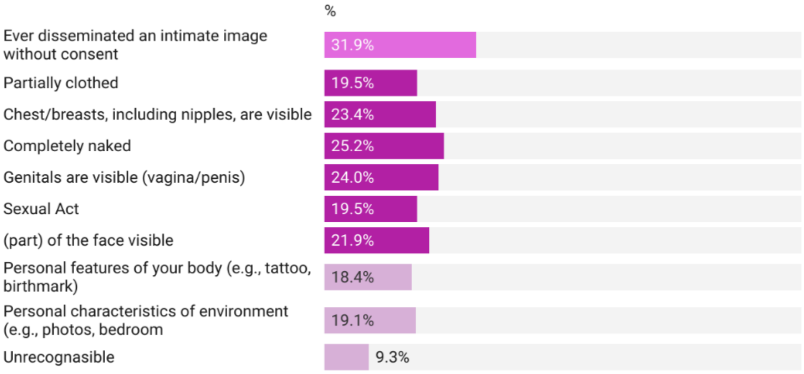


Figure 13. Content of the intimate image and victims' identifiers visible on the image

To disseminate the intimate image of someone else, one needs to save it first. One fifth of the perpetrators indicated that they had downloaded the image (20.6%) or saved the image in a cloud (20.5%). One fourth of the respondents indicated that they took a screenshot of the image (26.9%) or does not remember how they retrieved the picture. A smaller group of perpetrators printed the image (11%) or saved the image on an USB or hard drive (16.9%).

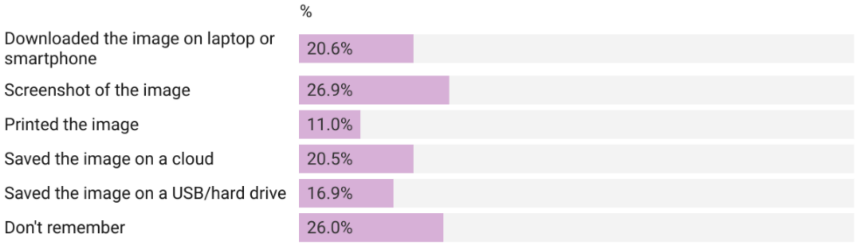


Figure 14. Method of saving the intimate picture

The aim of the study was to further investigate the role of gender, sexual orientation, ethnicity, and age in the context of NCII perpetration.

The role of gender

Table XIII (annex 11) shows that more than one third of male respondents are perpetrators of NCII (41.1%), whilst one out five women (22.7%) ever disseminated an intimate picture of someone else. One out of 10 of the subgroup "other" has ever disseminated an intimate picture of someone else (11.1%). The Chi-square test shows that the difference between women, men and the subgroup "others" is statistically significant (<0.01) and as such almost twice as many men disseminated intimate images of someone else than women.

The role of sexual orientation

Table XIV presents the prevalence differences in NCII perpetration according to sexual orientation. The results show that more than one-third of the LGBTQIA+ respondents are perpetrators of NCII (34.1%) whilst the prevalence rate of NCII perpetration is slightly lower in heterosexuals (30.8%). However, the Chi-square test shows that this difference based on sexual orientation is not statistically significant (p>0.05).

The role of ethnicity

Table XVI presents the prevalence differences in NCII perpetration according to the ethnic background of the respondent. Results show that almost half of non-Belgians with foreign background (42%) disseminated an intimate picture. Almost one out of three Belgian respondents (29%) and one fourth of Belgians with a foreign background (25.5%) has ever disseminated an intimate picture. However, the Chi-square test shows that this difference based on ethnicity is statistically significant (p<0.001).

The role of age

Table XVII shows that more than one third of the emerging adults (> 18 years old) has ever disseminated an intimate image without the consent of the depicted person. Moreover, results show that emerging adults (34.6%) disseminated more intimate images than adolescents (19.2%). The Chi-square test shows that this difference in NCII perpetration according to age is statistically significant (p<0.001).

Short summary on the role of diversity variables in NCII perpetration

In the above figures, some analyses show significant differences between diversity groups. There are no significant differences in NCII perpetration in terms of sexual orientation. There is a significant difference in terms of gender: almost twice as many men, and especially of the subgroup "other", disseminated an intimate picture. Non-Belgians and Belgians with a foreign background disseminated significantly more intimate images of someone else than Belgians. In terms of age, emerging adults spread significantly more intimate images of someone else in comparison to adolescents.

vi. How many people witness the non-consensual dissemination of intimate images?

Before perpetrators decide to disseminate the picture, they must retrieve the picture from somewhere else. By analysing the methods that perpetrators use to access intimate pictures, the results show that 22.7% of the perpetrators initially received the picture of someone else and thus became a bystander before disseminating it. 16.9% of the perpetrators saw the intimate picture on social media and/or the internet.

vii. Prevalence rates on harmfulness in vignettes of online hate speech

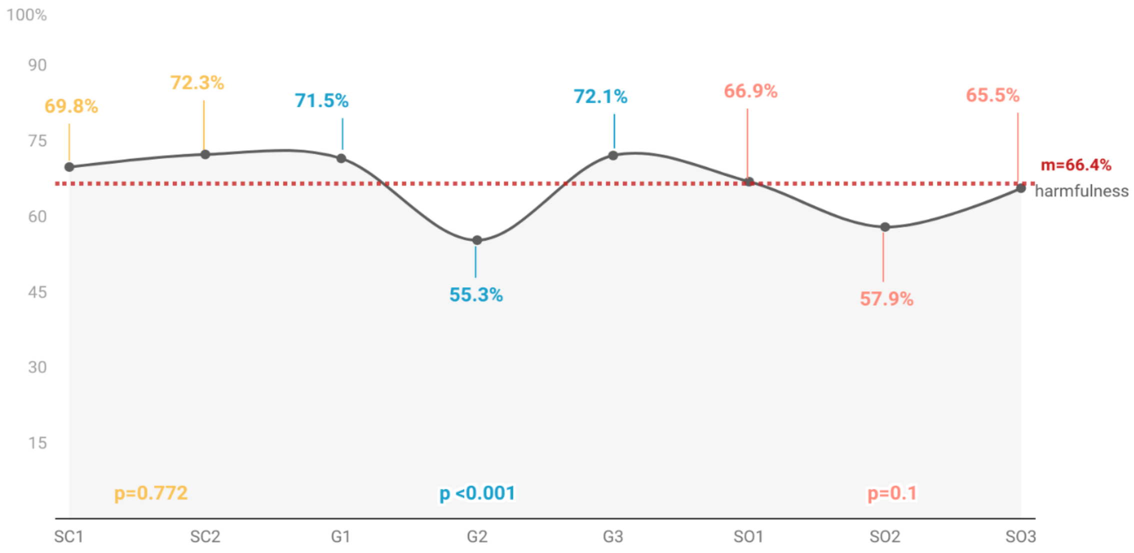


Figure 15. Harmfulness in vignettes of OHS

The results in figure 15 show that the average harmfulness in all vignettes concerning online hate speech is 66.4%. In other words, two-thirds of the respondents considered the described situations as harmful for the victim. Two vignettes, G2 and SO2, clearly are considered as less harmful by the respondents than all the other vignettes. In the vignette "G2" both the perpetrator and victim of OHS were male, whilst in vignette G1 the victim was female and the perpetrator male, and in G3 the victim was non-binary and the perpetrator male. Moreover, chi-square tests (table XXVIII, annex 11) revealed that there is a significant difference in the OHS vignettes that manipulate the victims' gender in terms of harmfulness: only 50% of the respondents think that the vignette G2 is less harmful than vignettes G1 and G3 (both more 70%). As such, it can be concluded that if the victim is male in OHS, adolescents and emerging adults think it is less harmful.

The vignette SO pictures a heterosexual victim that has been targeted for online hate speech, which might suggest that if the victim belongs to the LGBTQIA+ community, respondents consider this as more harmful. However, the chi-square tests show that there is no significant difference between these vignettes. There are no significant differences found in all vignettes of OHS regarding gender, sexual orientation, age, and ethnicity. Therefore, they are not presented in annex 11.

viii. Prevalence rates of harmfulness in NCII vignettes

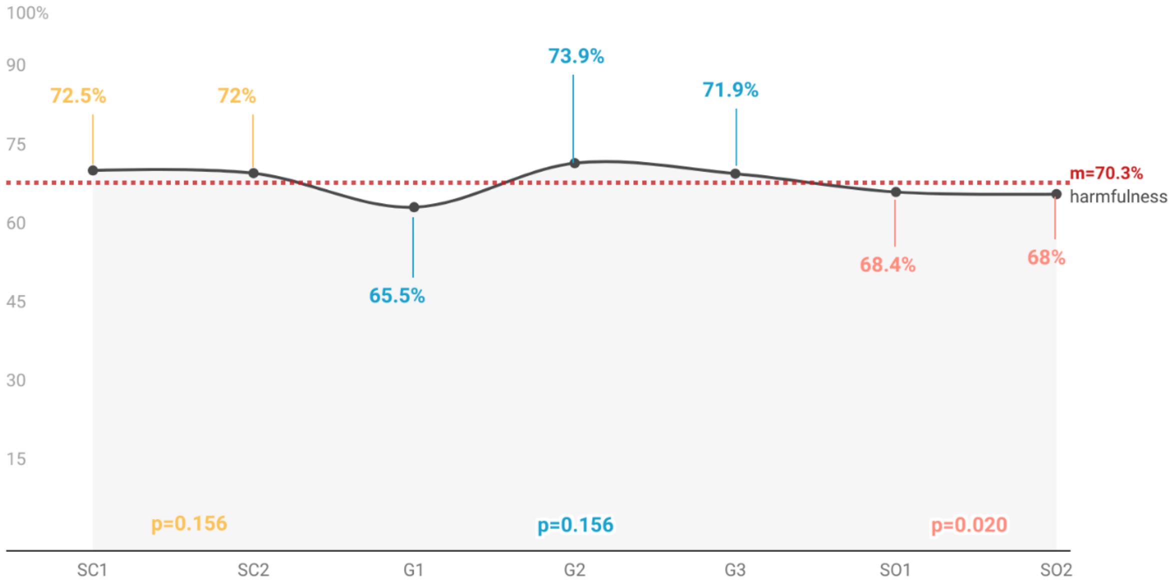


Figure 16. Harmfulness in vignettes of NCII

The results above demonstrated that, on average, 70.3% of the respondents assess the described NCII situations as harmful. For vignette G1, in which NCII victim is heterosexual, it appears that respondents find this behaviour the least harmful (65.5%) of all NCII vignettes. This difference was not statistically significant. Chi-square tests reveal (table XIX, annex 11) that there is a significant difference (p=0.020) in harmfulness between vignette SO1 and SO2. The victim of SO1 is heterosexual whilst the victim of SO2 is homosexual (in both vignettes the perpetrator is female) and as such respondents find NCII significantly more harmful when the victim is heterosexual as this is the only variable manipulated in the vignettes. In the vignette SO1 (victim is heterosexual), the respondent's ethnicity plays a significant role. More non-Belgians with a foreign background think that this vignette is harmful than Belgians and Belgians with a foreign background do. However, Belgians still find this vignette more harmful than Belgians with a foreign background.

There is also a significant difference between gender groups in the evaluation of harmfulness in vignette SC1. In this vignette, the ethnic background of the victim was manipulated (i.e., having a white skin colour) (table XXI, annex 11). Apparently, males find it significantly more harmful (76.4%) than women (69.3) when the person targeted for NCII has a white skin (table XXII, annex 11).

b. Findings on perspectives on legal action

i. Perspectives of adolescents and emerging adults on the types of legal action in OHS vignettes

For each vignette of OHS, respondents were asked which legal action they would like to be put in place within the context of OHS. Figure 17 shows that the majority of the respondents would like the perpetrator to follow a course about online violence (49.2%). One third of the respondents would like to imply a fine, community service or pay a damage compensation to the victim. Only one out of ten respondents would suggest an imprisonment.

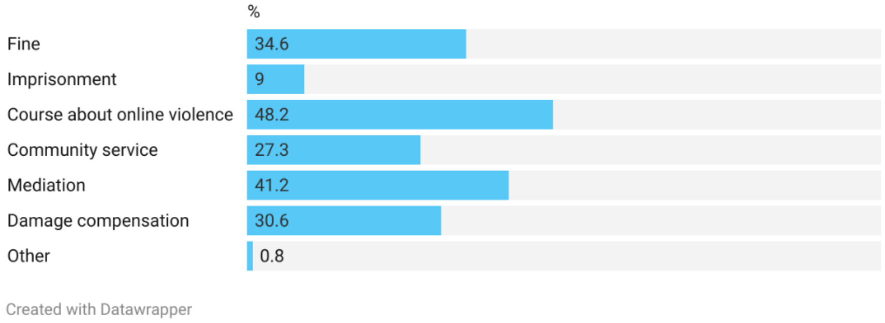


Figure 17. Prevalence of preferred legal actions by digital natives in OHS

Chi-square tests (table XXII to XV, annex 11) were done to analyse if there are significant differences (p≤ 0.05) between the subgroups of each variable. This section only included the differences that were significantly relevant. First, it was analysed if there were any significant differences in what legal action respondents would suggest for online hate speech, regardless of the types of vignettes. Across all the OHS vignettes, there were differences in the subgroups of sexual orientation and age whether to suggest a course on online violence. There were also significant differences in the subgroups of age whether to suggest a fine after OHS perpetration. There were significant differences in the subgroups of age whether to suggest imprisonment after OHS perpetration.

More adolescents than emerging adolescents think that a course on online sexual violence is an appropriate intervention. More adolescents find a course on online sexual violence appropriate for the perpetrator. Age also plays a significant role in how much they think a fine and imprisonment is an appropriate legal action. Emerging adults find a fine a more appropriate legal action after OHS perpetration, whilst more adolescents than emerging adults think that imprisonment is an appropriate method.

ii. Perspectives of adolescents and emerging adults on the types of legal action in NCII vignettes

For each vignette of NCII, respondents were asked which legal action they would like to be put in place within the context of NCII. Figure 18 shows that almost half of the respondents would like the perpetrator to pay a damage compensation to the victim (54.1%), to follow a course about online violence (45.7%), or to pay a fine (48.2%). Around 40% of the respondents would suggest a community service or a mediation process. One third of the respondents would like to imply a fine, community service (39.1%) or a damage compensation (39%) to the victim. One out of five of the respondents would suggest imprisonment.

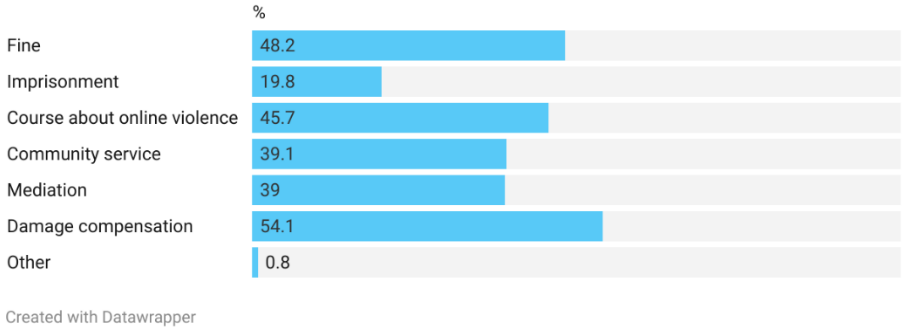


Figure 18. Prevalence of preferred legal actions by digital natives in NCII

Chi-square tests were done to analyse if there were significant differences between subgroups of the diversity variables (table XVI to XXXIII, annex 11). Overall, across all NCII vignettes, there is a significant difference between the subgroups of age on whether to suggest a course about online sexual violence, community service, pay damage compensation, and imprisonment. Sexual orientation plays a significant role on the perspective of applying for a course about online violence, community service, and a fine. There are significant differences between gender subgroups for suggesting imprisonment as appropriate after NCII perpetration.

More LGBTQIA+ members and emerging adolescents think that a course about online violence is appropriate after NCII perpetration in comparison to respectively heterosexuals and adolescents. The same is true for suggesting imprisonment; more LGBTQIA+ members (in comparison to heterosexual), females, and the subgroup "other" (in comparison to heterosexuals) indicate that this is an appropriate legal action. Adolescents and heterosexuals indicate that they find community service a more appropriate legal action in comparison to emerging adults and members of the LGBTQIA+ community. Regarding paying a damage compensation to the victim, more emerging adults indicate that they find this an appropriate legal action after NCII perpetration. More members of the LGBTQIA+ community indicated that they find implying a fine an appropriate legal action.

c. Testing PWM for OHS and NCII

i. Testing PWM for OHS

Measurement model

First, a measurement model (Model 1) was built to test whether the observed variables reliably reflected the hypothesised latent variables. The measurement model provided a good fit for the data: $\chi^2(125) = 484.615$, $p < .001$; CFI = 0.979, RMSEA = 0.036, CI [.032, .039], and SRMR = 0.032. All variables were treated as latent constructs, except for the single-item measure (i.e.,, prototype similarity and behaviour). All factor loadings were significant and above 0.44.

Structural equation model (SEM)

Secondly, we estimated a structural equation model (SEM) (Model 2) with intention and willingness as endogenous variables. The results of the structural model are presented in Figure 19. Results of the fit statistics indicated a good model fit: $\chi^2(159) = 741.76$, $p < .001$; CFI = 0.969, RMSEA = 0 .040, CI [.037, .043], and SRMR = 0.035. Our analyses revealed that attitude, subjective norm, and willingness explained 87.4% of the variance in intention, and that attitude, subjective norm, prototype similarity, and favourability explained 73.9% of the variance in willingness. The intention was significantly related to attitude (β = .15, p < .001), subjective norm (β = .13, p < .001), and willingness (β = .72, p < .001). Willingness was significantly related to prototype similarity (β = .36, p < .001) and prototype favorability (β = .60, p < .001). Behaviour was significantly related to intention (β = .40, p < .001), but not to willingness (β = .12, p =.11).

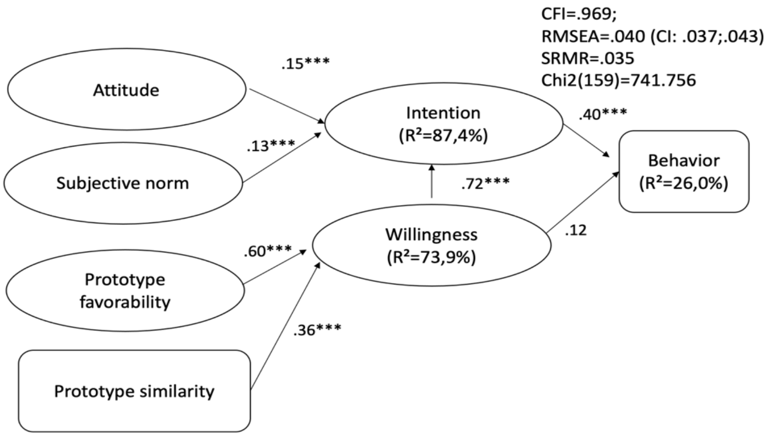


Figure 19. Prototype Willingness Model OHS

Thirdly, we tested a structural model with the following covariates: Age, Gender, Sexual orientation, ethnicity, prior victimisation, and financial stress. All covariates were associated with intention, willingness, and behaviour. Results of the fit statistics indicated a good model fit: $\chi^2(312) = 764.54$, p < .001; CFI =0 .958, RMSEA =0 .041, CI [.037, .045], and SRMR =0.059. Sexual orientation and ethnicity weren't correlated with any of the study variables. For reasons of parsimony, they were omitted from the analysis. The structural model with the three covariates (age, gender and financial stress) indicated a good model fit: $\chi^2(272) = 696.47$, p < .001; CFI =0 .960, RMSEA = 0.042, CI [.038, .046], and SRMR = 0.061.

Financial stress was significantly associated with behaviour (β = -0.13, p < .001), but not with willingness (β = 0.03, p = 0.21) and intention (β = 0.00, p =0 .98). Age was not significantly associated with intention (β = 0 .02, p = 0.10) but with behaviour (β = 0.09, p < .01) and willingness (β = 0.04, p = .03). Gender was significantly associated with willingness (β = -0.09, p < .001) and behaviour (β = -0.15, p < .001), but not with intention (β = -0.01, p = .40). Prior victimisation was significantly associated with willingness (β = 0.05, p < .01) and behaviour (β = 0.15, p < .001), but not with intention (β = -0.03, p = 0.14). The intention was significantly related to subjective norm (β = 0.18, p < .01), and willingness (β = 0.80, p < .001), but the association with attitude (β = 0.03, p = 0.57) disappeared. Willingness was significantly related to prototype similarity (β = 0.30, p < .001) and prototype favourability (β = 0.66, p < .001). Behaviour was significantly related to intention (β = 0.58, p < .001), but not to willingness (β = -0.02, p =.92).

To summarise, respondents' intention to engage in online hate speech is fueled by the subjective norms they experience concerning hate speech, in short, how significant others (e.g., parents and friends) would (dis)approve their involvement in hate speech. Moreover, prototype favourability and similarity were significantly related to willingness to engage in hate speech. This means that having a positive attitude towards a person who engages in online hate speech and perceiving oneself as being similar to that person, positively influences one's willingness to engage in hate speech. Furthermore, the model shows that engaging in hate speech is volitional rather than socially reactive, as behaviour was related to intention and not to willingness.

ii. Testing PWM for the non-consensual dissemination of intimate images

Measurement model

First, a measurement model was built to test whether the observed variables reliably reflected the hypothesised latent variables. The measurement model provided a good fit for the data: $\chi^2(84) = 290.98$, $p < .001$; CFI = .984, RMSEA = .033, CI [.029, .037], and SRMR = .028. All variables were treated as latent constructs, except for the single-item measure (i.e., prototype similarity, and behaviour). All factor loadings were significant and above .46.

Structural Equation Modelling (SEM)

Secondly, we estimated a structural equation model (SEM) with intention and willingness as endogenous variables. The results of the structural model are presented in Figure 20. Results of the fit statistics indicated a good model fit: $\chi^2(110) = 749.03$, $p < .001$; CFI = .953, RMSEA = .051, CI [.047, .054], and SRMR = .040. Our analyses revealed that attitude, subjective norm, and willingness explained 66.0% of the variance in intention, and that prototype similarity and favourability explained 83.2% of the variance in willingness. The intention was significantly associated with subjective norm (β = .21, p < .001) and willingness (β = .55, p < .001), but was not related to attitude (β = .08, p = 188). Willingness was significantly related to prototype similarity (β = .35, p < .001) and prototype favourability (β = .66, p < .001).

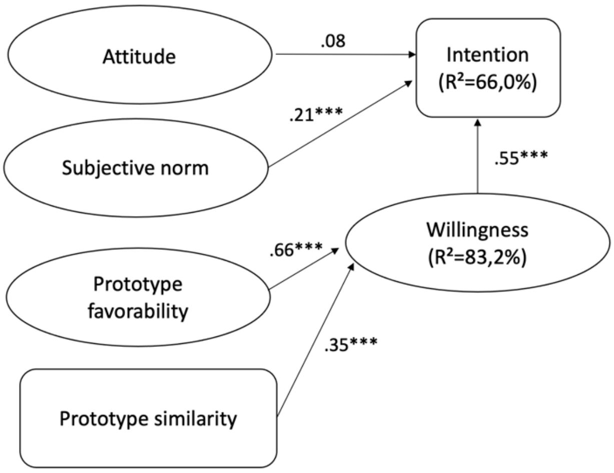


Figure 20. Prototype Willingness Model NCII

Thirdly, we tested a structural model with the following covariates: Age, gender, and sexual orientation. Results of the fit statistics indicated a good model fit: $\chi^2(155) = 964.83$, $p < .001$; CFI = .953, RMSEA = .049, CI [.046, .052], and SRMR = .065. Gender was significantly and positively associated with intention (β = -.09, p < .01) and willingness (β = -.10, p < .01).

The intention was significantly related to subjective norm (β = .23, p < .001) and willingness (β = .57, p < .001), but not to attitude (β = .06, p = .35). Willingness was significantly related to prototype similarity (β = .38, p < .001) and prototype favourability (β = .61, p < .001).

In sum, intention to engage in NCII is fuelled both by the reasoned path (but only subjective norm) and the social reactive path (i.e., willingness). Moreover, young people's willingness to engage in NCII is fuelled by the social reactive path (i.e., prototype favourability and similarity).

3.4 OSPS' SELF-REGULATORY FRAMEWORK & UNDERSTANDING OF CYBERVIOLENCE

3.4.1. METHODOLOGY

a. Research objectives and methodology

The research project aimed to get a better understanding of OSPs' self-regulatory framework, its interaction with the legal framework applicable to OSPs, and the way OSPs delineate online behaviour as (im)permissible, both in theory and in practice. This also entails an analysis of the proactive and reactive mechanisms applied to prevent, detect or remove impermissible content, as well as a survey on moderators' assessment of content as (im)permissible. To achieve these objectives, the research on the self-regulatory framework and understanding of cyberviolence by the industry encompasses different parts, namely (i) a mapping of the legal framework on OSPs and liability for illegal content, (ii) an analysis of OSPs' perspectives on cyberviolence, (iii) a review of the self-regulatory framework of OSPs and their delineation of permissible and non-permissible content with particular focus on OHS and NCII, and (iv) a study of the technical solutions to prevent and detect impermissible content.

The team relied on several methodologies relevant to the research objectives. First, the mapping of the legal framework on OSPs' role in combatting and their liability for illegal content was based on a classic legal analysis of relevant EU legal documents. This analysis became all the more important during the project with the entry into force of the new EU Digital Services Act (DSA), which was published in the Official Journal on 27 October 2022. Second, the analysis of OSPs' perspectives on cyberviolence was based on the input provided by representatives of relevant OSPs and industry associations during a roundtable stakeholder meeting organised by the team. Third, to better understand OSPs' delineation of (im)permissible content, the team performed an extensive literature study, coded and analysed the terms and conditions (T&C) and policy rules of a selection of OSPs, and conducted a survey among online content moderators. Fourth, the technical solutions deployed by OSPs to prevent and detect impermissible content were examined on the basis of the above coding of the T&C and policy rules.

b. Preliminary observations on terminology and scope

The team's discussions with OSPs revealed that the term 'Internet service providers', used in the initial @ntidote research proposal, no longer corresponds to the preferred terminology by these stakeholders. Moreover, the term 'Internet service providers' is not used in the EU, COE or in the Belgian national legal framework either, where other, more specific and sometimes partially overlapping terms can be found. Therefore, the team decided to search a more general term that is not tied to specific services or legislation but is broad enough to encompass online platforms and social media services used by digital natives. In consultation with industry, preference was given to the more inclusive term 'online service providers' (OSPs), which encompasses a broad range of services in the digital environment, including social networks, content storage services, communication services, online games and even streaming services. This term is also common among European LEAs (e.g., Sirius report 2022).

Second, it is important to highlight from the outset that the research into the self-regulatory framework of OSPs looks into the phenomenon of cyberviolence in a more general way, without focusing exclusively on OHS and NCII. There are various justifications for this methodological choice: 1) the term cyberviolence is not used by OSPs (see below, 3.4.2., e.); 2) the applicable EU legal framework exists independently of the category of cyberviolence concerned (with the exception of the Code of Conduct, see below, 3.4.2., a., iv.); and 3) the OSPs' self-regulatory framework applies more broadly to (im)permissible online content, which goes well beyond the two categories of cyberviolence that are at the heart of this project.

This said, the research in WP4 pays particular attention to OHS and NCII whenever this is possible or warranted.

3.4.2. RESULTS

a. Analysis of the normative framework of OSPs' role and liability

i. Introduction

OSPs' role in preventing and combatting impermissible or illegal content is not only defined by their own policies, it is also legally regulated, especially at EU level. The following analysis is exclusively dedicated to the EU legal framework, excluding the national legal framework. Several reasons explain this choice. First, the OSPs selected for this research (see below, c., i.) are active in Europe, and even worldwide, not just in Belgium. The analysis of their policies (see below, c., i.) shows that their self-regulatory framework is conceived globally, usually without considering national law. If there is any influence of local law on the T&C and policy rules adopted by the OSPs, it results solely from the EU legal framework. Second, with the entry into force of the DSA, which is directly applicable in the internal legal orders, MSs' domestic law concerning OSPs' obligations, including Belgian law (primarily contained in the Economic Law Code – art. XII.17 to XII.20), will no longer be applicable. Third, for the selection of OSPs, it was decided to concentrate on the EU market too as all platforms are available throughout the EU (or even globally).

At present, the EU legal framework on OSPs' role and liability consists of three instruments whose main elements will be examined below.

ii. e-Commerce Directive

Directive 2000/31/EC of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce in the Internal Market (commonly called the 'e-Commerce Directive'), is the first relevant EU text to consider the role of service providers in combating cyberviolence. It seeks to ensure the free movement of information society services between MS (art. 1, § 1), to eliminate legal obstacles arising from divergences in national legislation, and to ensure legal certainty as to which national rules apply to such services (recitals 5 to 8). From the outset, two points should however be emphasised. First, the Directive does not refer to 'cyberviolence' but only to 'illegal activity' or 'illegal information'. Second, the Directive's legal provisions on liability for online content (art. 12-15) are replaced by the DSA (especially by art. 4, 5, 6 and 8). When the DSA will become applicable to all providers included in its scope (17 February 2024), the Directive's liability provisions will no longer be applicable. But considering the latter applied to all OSPs at the moment of the analysis of their self-regulatory framework (April 2023) and continues to apply to many providers today, it is still useful to take a brief look at the Directive's scope of application and its liability exemption.

(1) Scope of application

This Directive applies to 'information society services providers' (on this term, see CJEU 11 September 2014, C-291/13) established on the territory of a MS (art. 3, §1). This comprises several subcategories of providers: providers of 'mere conduit', of caching services and of hosting services (art. 12-14). The term 'information society services' covers any service normally provided for remuneration, at a distance, by electronic means and at the individual request of a recipient of a service (recitals 17-18).

(2) No monitoring obligation and exemption from liability for third-party content

Importantly, the e-Commerce Directive entails an exemption from liability for third-party content and does not require the service providers to monitor information provided by third parties. Quite the contrary, the Directive even prevents MS from imposing a general obligation on providers to monitor the information they transmit or store, or to actively seek facts or circumstances indicating illegal activity (art. 15, § 1 & recital 47). The liability exemption depends on several conditions, which vary according to the three types of services distinguished by the Directive: mere conduit, caching, and hosting (for a detailed overview, see annex 12).

MSs may nevertheless require providers of hosting services to apply a duty of care to detect and prevent certain types of illegal activities (recital 48), as the CJEU clarified in its decision *Glawischnig-Piesczek v. Facebook Ireland* (3 October 2019, C-18/18). Furthermore, MSs can establish obligations for providers to promptly inform the competent public authorities of alleged illegal activities undertaken or information provided by recipients of their service, or obligations to communicate to the competent authorities, at their request, information enabling the identification of recipients of their services with whom they have storage agreements (art. 15, § 2). Neither does the liability exemption affect the possibility of injunctions of a court or administrative authority requiring the termination or prevention of any infringement (art. 12, § 3; art. 13, § 2; art. 14, § 3), the removal of illegal information, or the disabling of access to it (art. 14, § 3).

iii. The Digital Services Act (DSA)

(1) General objectives of the DSA

The DSA aims to contribute to the proper functioning of the internal market for intermediary services by setting out harmonised rules for a safe, predictable, and trusted online environment that facilitates innovation and in which fundamental rights enshrined in the Charter, including the principle of consumer protection, are effectively protected (art. 1, §1). Like its predecessor, the e-Commerce directive, the DSA entails a conditional liability exemption for providers of intermediary services (art. 1, §2, a). Yet, it innovates by adding a comprehensive list of new due diligence obligations tailored to certain specific categories of providers of intermediary services, to reduce harm and counter risks online (art. 1, §2, b)) (overview in annex 13).

The DSA entered into force on 16 November 2022, but will only gradually become applicable (for more details, see Figure 21). At the time of writing, the DSA already applies to very large online platforms (VLOPs), including several OSPs selected for the analysis of the self-regulatory framework (see below, c., i.), and it will be directly applicable across the EU to all services providers within its scope from 17 February 2024 (art. 93, §2).

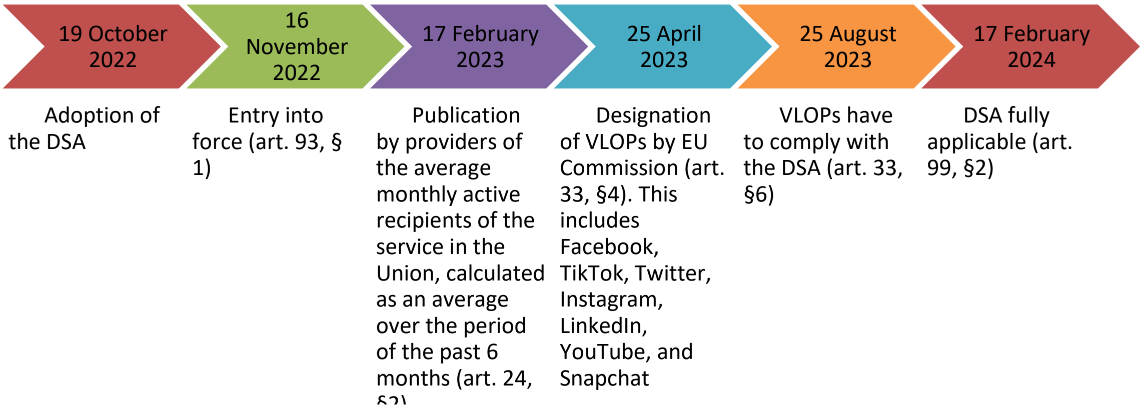

19 October 2022
Adoption of the DSA
16 November 2022
Entry into force (art. 93, § 1)
17 February 2023
Publication by providers of the average monthly active recipients of the service in the Union, calculated as an average over the period of the past 6 months (art. 24, §2)
25 April 2023
Designation of VLOPs by EU Commission (art. 33, §4). This includes Facebook, TikTok, Twitter, Instagram, LinkedIn, YouTube, and Snapchat
25 August 2023
VLOPs have to comply with the DSA (art. 33, §6)
17 February 2024
DSA fully applicable (art. 99, §2)

Figure 21. Chronology of DSA's applicability

(2) Personal and territorial scope of application

The DSA applies to providers of 'intermediary services' (art. 2, §1) defined by reference to three categories of services, i.e., mere conduit, caching and hosting services – the same, with a very similar definition, as in the e-Commerce Directive. The DSA, however, introduces a new criterion to define the territorial scope. The DSA applies to services that are offered to recipients who have their place of establishment or are located in the Union, irrespective of the place of establishment of the providers of those services (art. 2, §1). Unlike the e-Commerce Directive, the DSA's territorial scope thus targets the location of the services' recipient, rather than the location of the service provider.

(3) Exemption from liability for third-party content

Just like the e-Commerce directive, the DSA exempts providers of intermediary services from liability for third-party content (art. 4-6). It however specifies that this exemption should apply in respect of any type of liability (civil, criminal, or contractual), regarding any type of illegal content, and irrespective of the precise subject matter or nature of the relevant laws (recital 17). The liability exemption is subject to the same conditions that vary depending on the type of service provided as under the e-Commerce Directive (see art. 4-5-6 DSA and the detailed overview in annex 12).

Importantly, the DSA specifies that providers of intermediary services shall not be deemed ineligible for the liability exemptions solely because they carry out, in good faith and in a diligent manner, voluntary own-initiative investigations into, or take other measures aimed at detecting, identifying and removing, or disabling access to, illegal content, or take the necessary measures to comply with the requirements of Union law and national law in compliance with Union law, including the requirements set out in the DSA (art. 7). For instance, the fact that the provider takes steps to ensure compliance with its T&C will not automatically lead to the conclusion that it should have been aware of the illegal content and thus should have reacted expeditiously to remove that content.

The situation is however different when hosting services providers, including online platforms, receive notice related to content that is to be considered illegal content: if the notice contains sufficient information to enable a diligent provider to identify, without a detailed legal examination, that the content is illegal, the notice should be considered to trigger actual knowledge or awareness of illegality, giving rise to liability if the providers does not act expeditiously to remove or to disable access to the information (see art. 16, § 3 & recital 53).

(4) No definition of illegal content

It is important to stress that the DSA does not contain any precise definition of what constitutes illegal content. Indeed, it does not harmonise what content or behaviour counts as illegal. It only indicates that illegal content means 'any information, which, in itself or in relation to an activity, including the sale of products or the provision of services, is not in compliance with Union law or the law of any Member State, irrespective of the precise subject matter or nature of that law' (art. 3, h). For further definition, it refers to EU law and MS law; some EU instruments indeed contain such definition (e.g., CSAM Directive and Framework Decision on racism and xenophobia). That said, it is noteworthy that the DSA uses the term 'cyber violence' (recital 87) by reference to certain forms of illegal content, including illegal pornographic content, and with respect to victims of non-consensual sharing of intimate or manipulated material.

(5) Due diligence obligations for a transparent and safe online environment

Undoubtedly, the true innovation of the DSA lies in the new due diligence obligations for providers of intermediary services. They seek to ensure a safe, transparent, and predictable online ecosystem. Those obligations are not linked to the liability exemption. The failure to comply with them can only result in penalties for the providers in accordance with article 52. Furthermore, those obligations continue to apply even when providers fail to meet the conditions of the liability exemption. The liability of providers of intermediary services must therefore be assessed separately (recital 41).

Despite these new obligations, it is important to stress that, like the e-Commerce Directive, the DSA does not impose a universal obligation to moderate content (art. 8). There is even a prohibition to impose on providers a general obligation to monitor information or to actively seek facts or circumstances indicating illegal activity (art. 8), to the extent previously mentioned by the CJEU in its decision *Glawischnig-Piesczek v. Facebook Ireland* (3 October 2019, C-18/18).

Nevertheless, since the DSA contains specific obligations for providers with regard to content moderation, it also contains a definition of content moderation: 'activities, whether automated or not, undertaken by providers of intermediary services, that are aimed, in particular, at detecting, identifying and addressing illegal content or information incompatible with their terms and conditions, provided by recipients of the service, including measures taken that affect the availability, visibility, and accessibility of that illegal content or that information, such as demotion, demonetisation, disabling of access to, or removal thereof, or measures that affect the ability of the recipients of the service to provide that information, such as the termination or suspension of a recipient's account' (art. 3). Content moderation is thus defined broadly, encompassing both proactive and reactive content moderation mechanisms.

The due diligence obligations in the DSA vary according to the type, size, and nature of the intermediary service. The combination of these criteria leads to a distinction between five categories of service providers relevant for the @ntidote project:

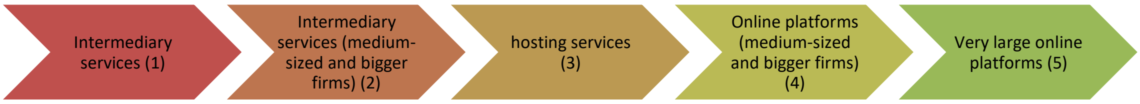

Intermediary services (1)
Intermediary services (medium-sized and bigger firms) (2)
hosting services (3)
Online platforms (medium-sized and bigger firms) (4)
Very large online platforms (5)

Figure 22. Categories of providers regarding due diligence obligations under the DSA

(1) All providers of intermediary services, in accordance with the DSA definition described above (mere conduit, catching and hosting).
(2) Medium-sized or bigger firms (as defined in Commission Recommendation 2003/361/EC, annex, art. 2) providing intermediary services, i.e., firms with 50 employees and more or with more than 10 million EUR turnover or annual balance sheet, with the particularity that a micro or small-sized enterprise which becomes a medium-sized or bigger firm has to meet the related obligations 12 months following the change of status, except when they become a VLOP (art. 19, §1).
(3) Providers of hosting services, regardless of the size of the enterprise (art. 3, g), iii).
(4) Online platforms, i.e., platforms hosting third-party information but also disseminating it to the public as a core feature (art. 3, i)) and which constitute medium-sized or bigger firms (art. 19, §1).
(5) VLOPs, i.e., online platforms which have a number of average monthly active recipients of the service in the Union equal to or higher than 45 million (equivalent to 10% of the Union population) and which have been designated as such by the European Commission, regardless of the size of the firm that provided such platforms (art. 33, §1). An 'active recipient of a service' is engaging with the service at least once in a month, by being exposed to the information disseminated on the online interface of the platforms, such as viewing it or listening to it or by providing information without limitation to interacting with information by clicking on, commenting, linking, sharing, purchasing, or carrying out transactions on an online platform. It is not necessarily a registered user (recital 77).

To the extent that providers of intermediary services belong to different categories (e.g., in view of the nature of their services and their size), they should obviously comply with all the corresponding obligations of the DSA (recital 41 – see annex 13 for a schematic overview).

(6) Obligation to cooperate with public authorities

The DSA imposes obligations on service providers to cooperate with national public authorities. Still, there is no obligation to act against illegal content following an order issued by a national authority, only an obligation to inform the authority issuing the order or any other authority specified in the order of any follow-up given to the orders, without undue delay, specifying if and when the order was applied (art. 9, § 1). This may be regrettable, but the choice made by the EU legislator is also understandable, considering the rule of law and fundamental rights issues in certain MSs. Moreover, the DSA has the merit of harmonizing certain minimum conditions that the injunction must meet (e.g., language used, information to include in the order such as legal basis, statement of reasons explaining why the information is illegal content) (art. 9, §2). It should be noted though, that the DSA does not provide a legal basis for such orders, nor does it regulate their territorial scope or cross-border enforcement; this is left up to national law.

The DSA does, however, state that the territorial scope of the order must be limited to what is strictly necessary to achieve its objective (art. 9, §2, b). In a cross-border context, the effect of the order should in principle be limited to the territory of the issuing MS, unless the illegality of the content derives directly from EU law, or the issuing authority considers that the rights at stake require a wider territorial scope (recital 36).
The enforcement of an order to act against illegal content is also governed by national law, in compliance with Union law, including the Charter and the TFEU provisions on the freedom of establishment and the freedom to provide services within the EU (recital 32).

In addition, national authorities can issue a second kind of order to obtain information about the recipient of a service (art. 10). The DSA harmonises the conditions the order must meet, especially the information it must contain (art. 10, §2), e.g., the fact that the recipient must be clearly identified in the order (recital 37). Other orders, targeting a group of persons or seeking aggregate information required for statistical purposes or evidence-based policy making, are not covered by article 10 (recital 37). Again, there is no obligation to act but only to inform the relevant authority if and when effect was given to the order (art. 10, §1). This is, however, without prejudice to Regulation (EU) 2023/1543 of 12 July 2023 on European production orders and European preservation orders for electronic evidence in criminal proceedings and for the execution of custodial sentences following criminal proceedings, which will impose (as of 18 August 2026) an obligation of cooperation on service providers when they are asked to provide information, including subscriber information, but only during an ongoing criminal investigation.

iv. The Code of Conduct on countering illegal hate speech online

In addition to the above two legal instruments, there is one soft law instrument that is highly relevant for the @ntidote project, and especially OHS: the Code of Conduct on countering illegal hate speech online. The Code of conduct has been signed by Facebook, Microsoft, Twitter, YouTube, Instagram, Snapchat, Dailymotion, Jeuxvideo.com, TikTok, LinkedIn, Rakuten Viber and Twitch. Preceding the DSA, the Code constitutes an effort to respond to the challenge of ensuring that online platforms do not offer opportunities for illegal online hate speech to spread virally. It is aimed at guiding signatories' activities as well as sharing best practices with other internet companies, platforms, and social media operators.

The Code's scope of application is however limited: it centres on hate speech which is defined, unsurprisingly, by reference to the definition of illegal hate speech under the Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law, e.g., 'all conduct publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin' (art. 1, §1, a)).

The Code of Conduct already contained several commitments that are now to be found, at least in part, in the DSA with a more general scope, including the commitment to:

- Have clear and effective processes to review notifications regarding illegal hate speech, through a dedicated team, against the provider's rules and community guidelines and, where necessary, national laws transposing the 2008 Framework Decision on combating certain forms and expressions of racism and xenophobia by means of criminal law;
- Review the requests for removal in less than 24 hours and, if necessary, remove or disable access to such content;
- Raise awareness with the users about (im)permissible content, including through the notification system;
- Provide information on the procedures for submitting notices, in order to improve the swift and effective communication between MSs and signatories on notifications and on disabling access to or removal of illegal hate speech online, especially through national contact points designed by the signatories;
- Provide regular training to their staff on current societal developments and exchange views on the potential for further improvement;
- Cooperate with other signatories of the Code as well as other platforms and social media companies to enhance best practice sharing;
- Cooperate with 'trusted reporters' or 'trusted flaggers', i.e., civil society organisations, particularly by transmitting information and providing support and training on the company rules, the community guidelines, and rules on the reporting and notification processes, while taking care to

respect their independence and credibility. The objective is also to engage with trusted flaggers on a regular basis to increase understanding of national specificities of hate speech. Signatories are supposed to make information about 'trusted reporters' available on their websites;

- Provide transparency toward users (concerning the outcome of flagging and feedback on the decision regarding the posted content) and the public (via transparency reports on the enforcement of community rules).

Participation in the Code is voluntary, but once a company has signed up, it is subject to monitoring rounds and must report annually on how it counters online hate speech. The results of annual monitoring on compliance with the commitments contained in the Code of Conduct are made public and are available on the website of the European Commission. To date, the last monitoring round was carried out from 28 March to 13 May 2022, but did not yet include Twitch, Rakuten Viber and LinkedIn. The results were published in November 2022. A few elements of interest to the @ntidote project can be highlighted:

- The number of notifications reviewed within 24 hours (64.4%) has decreased as compared to 2021 (81%) and 2020 (90.4%). Only TikTok has increased its performance (from 82.5% to 91.7%);
- The average removal rate (63.6%) is similar to 2021 (62.5%), but still lower than in 2020 (71%) and most of the signatories (except YouTube) have removed less hate speech content than in 2021;
- The quantity of feedback to user notifications has improved as compared to previous monitoring exercise (66.4% of notification received, to 60.3% in 2021), but this feedback is more widely spread to trusted flaggers than to general users;
- Antigypsyism (16.8%), xenophobia including anti-migrant hatred (16.3%), and sexual orientation (15.5%) are the most reported grounds of hate speech during the monitoring exercise in 2022, but this data is only an indication, as it is influenced by the number of notifications sent by trusted flaggers as well as the delineation of their work field.

The report also entails the rate of removals per country, depending on the country of origin of the trusted flaggers and on the number of notifications sent by each of them. During the monitored period, there were no notifications from Belgium.

b. OSPs' perspectives of cyberviolence

Within the ambit of the project, a roundtable stakeholder meeting was organised with representatives of OSPs and industry associations, i.e., professional organisations representing the industry (annexes 14 and 15).

The OSPs invited to take part in the roundtable were selected based on what the participants to the interviews in the qualitative research on the understanding of cyberviolence indicated as frequently used online platforms (see above 3.3). This selection was broader than the subsequent selection made for the analysis of the self-regulatory framework (see below, c., i.), which can be justified by the fact that the roundtable was intended to be an exploratory discussion. Participation in the meeting was of course voluntary. Not all invited OSPs responded to the invitation. When an OSP indicated that it could not attend on the scheduled date, the team proposed a one-by-one meeting instead. Several industry associations were also contacted and responded positively. Eventually, the roundtable gathered two industry associations and five representatives of widely used OSPs. The roundtable was conducted in accordance with Chatham House Rules.

The OPSs taking part in the roundtable belong to both first and second (i.e., more recently established) generation of online platforms; some are used almost exclusively by digital natives and others more widely by both digital natives and adults. The roundtable also included a representative of a platform dedicated to pornographic content. In addition to the roundtable, the team was able to organise separate meetings with representatives of two other OSPs, covering the same topics as the ones selected for the roundtable. All participating OSPs are globally active, not just in Belgium or the EU.

The topics for discussion with the industry focused on the definition and understanding of cyberviolence by OSPs, their perception of the role they play in preventing, detecting, and combating cyberviolence, and on the technological tools to address such behaviour (for the full list of topics and programme of the roundtable, see annexes 14 and 15).

Before looking into some findings, two general remarks should be made. First, even if the discussions with industry stakeholders have provided valuable input on certain research questions in a different way than by consulting and analysing the information published on their websites, it should be noted that the information the participants shared during the roundtable and one-on-one meetings largely corresponds to the T&C and policy rules available on their websites. It seems difficult for industry representatives to go beyond this publicly available information, which suggests that cyberviolence and content moderation are relatively sensitive subjects for OSPs. Second, while another objective was to establish contacts with OSPs, the willingness to cooperate in the moderators' survey proved to be limited. All participants were very interested in the research and offered their help, but when it came to obtaining more information on the implementation of the self-regulatory framework and operational aspects of online content moderation, most doors remained closed. This too confirms the sensitivity of the research topic.

i. OSPs' understanding of cyberviolence and assessment of (im)permissible content

A first important finding is that none of the OSPs use the term 'cyberviolence' in their T&C and policy rules. Some participants had apparently never heard the term before. Others underlined that the term was too broad. The participating OSPs indeed prefer to distinguish between certain types of cyberviolence and to use more specific terms (e.g., hate speech/hateful content, child sexual exploitation, abuse, bullying/harassment, nudity and sexual content, near sexual images, and threats) and to adopt separate policies for different forms of cyberviolence. This sequenced approach makes the broad phenomenon of cyberviolence more sizeable, enabling OPSs to react quickly when there is a need to adapt the policy rules.

Every OSP, and even every platform (even if controlled by one and the same OSP), has its own T&C and policy rules - this has also been confirmed by a more in-depth analysis of the self-regulatory framework (see below, c., i.). These terms and rules have been developed 'from scratch' and have evolved a lot over time. For OSPs established in the U.S., there are real tensions between what top-level company officials consider to be permissible content, as they are strongly committed to free speech, and what is considered problematic content by European norms, even if these OSPs have made progress to meet European standards. Some OSPs indicated their company rules have been elaborated or further adjusted in cooperation with experts, from both academia and civil society; others mentioned the use of surveys and adjustments based on court decisions. Some also explained that they try to find the common denominator in the various legal frameworks they are subject to when developing globally applicable policies.

This can result in a self-regulatory framework that is stricter than the legal framework of certain countries. Other participants also mentioned the adoption of 'global guidelines with local enforcements', depending on the origin of the posted content. Interestingly, several participants believed the DSA would change a lot for OSPs, including at the level of the content of their policies and the definition of (im)permissible content, as it provides for dialogue and a framework for new codes of conduct (in addition to the one on OHS, presented above).

The industry associations, for their part, explained how their members (especially smaller OSPs) cooperate with one another to develop policy rules on NCII and OHS, and draw inspiration from the policies and practices of other bigger OPSs. They highlighted the challenge to deal with different legal frameworks, since there is no common approach to what constitutes impermissible content.

Consequently, the categories or labels used in the self-regulatory framework differ from one OSP to another (as confirmed by the subsequent in-depth analysis of OSPs' policies) and are not immutable. On the contrary, the policy rules developed to deal with impermissible behaviours are constantly evolving. Whenever it is necessary to tackle a new kind of behaviour or phenomenon, the self-regulatory framework will be adapted and, if need be, a new category and specific policy will be developed. Sometimes policy rules are also amended to exclude certain content that would otherwise be considered impermissible; to this end, an exception is created, or even an exception to the exception. This process of very detailed sub-categorising policies was described by one participant as a 'ramification process'. This highly detailed approach to what is (im)permissible content corresponds to previously leaked documents on internal training and guidance given to content moderators (Gillespie, 2018; Hartwig & Heckenlively, 2021) as well as interviews with (former) moderators in big tech companies, describing internal policies as 'very, very specific itemized' rules (Roberts, 2021).

Interestingly, several participants emphasised there are no grey zones in company policies on content moderation, even if there may exist some in the implementation process. Content is either permissible or impermissible, either legal or illegal; every type of content necessarily goes into one 'bucket' or 'pocket'. Of course, mistakes are possible, but they are corrected thanks to the internal review process. For instance, AI tools may classify content as admissible ('fine'), inadmissible ('not fine') and 'grey zone'; the latter category then calls for a review by a human moderator. This strong assertion that policy rules are free of grey zones is quite remarkable because the research results of WP1, WP2 and WP3 clearly show that NCII and especially OHS are hard to define as real-world situations are full of grey zones. Moreover, to be able to deal with such grey zones, it is well-known that OSPs' publicly available policies (which, contrary to the internal rules, are written in open, 'nebulous language') leave 'wiggle room' for moderators (Roberts, 2021). The latter was also confirmed by the participants in the roundtable: policies allow for flexibility in assessment because they are designed to be applied globally, and they need to be changed on a regular basis to deal with new behaviours and changing conceptions of what is socially acceptable. Arguably, this leads to a paradox: on the one hand, OSPs and industry associations assert that there are no grey zones and that their policies are adjusted to all real-world situations, creating a reassuring image; on the other hand, they acknowledge that the policy rules are a living instrument, in need of frequent updates. This testifies to the difference between the publicly available policies and the internal rules designed to give clear guidance to moderators and to enable them to make quick decisions and avoid subjectivity. As one participant formulated it: 'Moderators are not based on subjectivity and are extremely consistent. Subtility is baked in the rules.' All in all, it underscores the importance of the self-regulatory framework for content moderation.

ii. Content moderation process and use of technological tools to address cyberviolence

All OSPs confirmed they have content moderation techniques in place, consisting of a mix of human and automated moderation. AI and other technological tools are used to address problematic behaviour; they include both reactive and proactive mechanisms to (automatically) detect and remove, or to quarantine problematic content until a human moderator reviews it. The need to combine the use of artificial and human intelligence was reiterated several times, as well as the fact that AI is a good (first) filter or classifier. For instance, AI can detect pornography and identify whether the person represented on the picture is (or might be) a child. All participants, however, agreed that AI has certain drawbacks and that certain cases are difficult to address by AI.

Nevertheless, it appeared difficult to have a frank and open discussion on the type of content moderation processes used, their technical operation, as well as their advantages and weaknesses. The input remained rather general. More precise information could not be shared, as these topics touch upon business secrets.

Only a couple of OSPs around the table were willing to explain some aspects of their content moderation process, such as the training provided to moderators, the use of local moderators with local language skills to reflect local values and dynamics in content moderation decisions; the fact that the content moderation tools implemented may vary depending on the type of content and whether it concerns images, videos or text; the possibility to review a piece of content twice in sensitive cases, like those involving politicians; and the possibility of limiting the territorial application of a content moderation decision to certain countries where the content is illegal.

iii. OSPs' perception of their role in preventing, detecting & combatting cyberviolence

Finally, all participating OSPs were aware of the important role they play in detecting and combatting impermissible content. This is demonstrated by the initiatives they have sometimes taken, such as their cooperation with NGOs, not only to flag illegal content but also to build and develop their policies or to raise awareness among their users (for instance, through short videos made by NGOs and widely distributed by the provider). The OSPs in the roundtable also emphasised the efforts made to prevent their users from being exposed to illegal content and, in general, to maintain a safe and healthy digital environment.

c. The self-regulatory framework of OSPs on permissible and non-permissible content

i. Analysis of the OSPs' T&C and policy rules

Selection of OSPs for study of self-regulatory framework

Before engaging in the analysis of the self-regulatory framework, the team had to select the OSPs to include in this analysis (annex 16). In the initial project proposal, the OSPs had been selected based on three criteria:

1) Activity in Belgium; and
2) Popularity among adolescents and young adults as monitored by the latest Mediawijs and Mediaraven bi-annual report 'Apestaartjaren'; and/or
3) Regular mentions by the participants of the interviews of the project on qualitative understanding (WP1) and of the vignette-study among digital natives (WP3).

The application of these criteria did, however, not yield a sufficiently specific and adequate selection. The team therefore further refined and completed those criteria to obtain a sufficiently representative and diverse selection, while ensuring the feasibility of the study. Eventually, the criteria used to select the OSPs and, more specifically, the platforms they operate, are the following ones:

1) The accessibility from the Belgian territory;
2) The popularity among digital natives, assessed not only on the ground of the last (2022) Mediawijs and Mediaraven bi-annual report 'Apestaartjaren', but also based on data collected from the survey carried out in WP3;
3) The type of service provided: the core business of the selected platforms consists of sharing information in public or with a large number of persons, thus excluding streaming platforms, purely private messaging platforms as well as gaming platforms;
4) The State in which the OSP running the platform has established its headquarters: this enabled us to include in our study platforms controlled not only by U.S.-based OSPs, but also by OSPs based in China, in Russia (originally), in Canada, and in France;
5) The 'age' of the platform to include both first (e.g., Facebook, Twitter, YouTube, LinkedIn) and second (e.g., TikTok, Discord, BeReal) generation platforms;
6) The participation of OSPs running the platform in the Code of Conduct on countering illegal OHS;
7) The type of content moderation mechanisms in place: proactive and/or reactive tools, the use of technical tools and/or human moderation, and the fact that moderation is carried out by a team of professional moderators and/or by users themselves;
8) The type of content hosted: preference was given to platforms used for sharing speech (including written speech and comments) and image-related content, including a pornographic platform, since NCII is one of two types of cyberviolence the @ntidote project focuses on.

Based on those eight criteria, twelve platforms were selected. In some cases, the platform name is identical to the OSP's name, but not for others. In addition, in a couple of cases, OPSs manage several platforms. In the rest of the analysis, we will continue to refer to OSPs, bearing in mind that this term may refer to specific platforms which will then be explicitly indicated. Further details of this selection and their main characteristics can be found in annex 16.

Assessment of the self-regulatory framework of selected OSPs

OSPs run not only the platforms on which content is posted and distributed (i.e., instruments), they also function as regulators in the sense that they determine the T&C for online content, decide on the technical possibilities for online actions (e.g., liking, dissemination), the algorithms pushing or suppressing certain content, and the actions taken when content is reported to constitute OHS or NCII. This situation revealed the importance of the assessment of the self-regulatory framework of the previously selected social medial platforms.

The self-regulatory framework consists, primarily, in the T&C (also called 'terms of use', 'terms of service', or 'user agreement' by certain OSPs) and the policy rules or community guidelines of the selected platforms. In addition, other relevant publicly available information on the websites of these OSPs was taken into account. All of this was downloaded in the same month (April 2023) and with indication of the date of last update of the T&C, in order to have a snapshot of the self-regulatory framework of all OSPs at the same point in time. This is important as those rules tend to change regularly, and even more so because the DSA, which was not yet applicable at the time of the analysis, is expected to considerably impact this self-regulatory framework.

Subsequently, all this information was analysed based on the 'coding technique', which is the same technique as the one used for the analysis of Belgian court files in part 3.2 on the legal framework of OHS and NCII. A first step consisted in the preparation of an analytical grid in the form of an Excel document. This grid comprises a whole range of elements relating to the categories of permissible/impermissible online content, the definition of NCII and OHS, the impact of the legal framework, the proactive and reactive mechanisms put in place by the OSPs to prevent, detect and delete impermissible content, and the moderation process, as well as the consideration by OSPs of their role vis-à-vis authorities through the commitment of reporting illegal content. While these elements clearly relate to the research objectives, some of these elements were added or refined in light of the stakeholder meeting with industry (see above, b.). Others have been inspired by the literature review conducted for the survey among moderators (see below, ii.). An overview of the analytical grid can be found in annex 17. Once the grid finalised, the collected data were analysed.

General findings

1) Accessibility of OSPs' policies

The T&C and policies of the selected platforms are adequately published and accessible online. There are, however, certain exceptions. For instance, Telegram's policies consist only of a FAQ, indicating that they continue to evolve and will be completed with new features in the next few months.

2) Segmented approach

The OSPs' self-regulatory framework has not been developed in a comprehensive way to prevent and combat all the forms of impermissible behaviour in the same way. On the contrary, the policies have generally been designed on a segmented basis. Each policy defines separate permissibility criteria, which vary according to the type of behaviour and the category of behaviours it belongs to. This corresponds to the findings of the roundtable stakeholder meeting (see above, b., i.).

OHS and NCII are covered, more or less extensively, by the policy rules, even if the terminology OSPs use varies considerably. For instance, the OSPs use the terms 'hateful behaviour', 'hate speech or symbols and encouragement of violence or attacks', 'harassment when it threatens or targets an individual based on intrinsic attributes and hate speech', and simply 'hate speech'. The terminology used to designate NCII is less specific and generally refers to other, broader behaviours. It refers, for instance, to non-consensual intimate activity and sexualisation of minors, nudity or sexual activity, sexual graphic objectification of an individual without its consent, as part of abuse/harassment policy and non-consensual nudity as part of the sensitive media category, gender-based violence and sexual exploitation including image-based sexual abuse, or even (only) revenge porn.

3) Room for discretion

As indicated, during the roundtable with industry, OSPs presented their publicly available policies as a living framework allowing for some flexibility. The analysis of the self-regulatory framework confirms this. On the one hand, the T&C and policies are regularly updated. In the selected sample of OSPs, the rules analysed had been in force for several months, with a maximum of two years. On the other hand, some OSPs provide very detailed rules to their users, while others are much more succinct. Certain OSPs even specifically dictate do's and don'ts to their users. However, they all contain a certain degree of flexibility, leaving room for discretion in the moderation process. This does not necessarily mean that there is a lot of room for flexibility internally. As explained below (ii.), the survey among moderators too shows that the internal rules leave very little room for interpretation by the human moderators.

4) Limited impact of the local legal framework

Generally, OSPs' policies define impermissible content without making a distinction based on the location of the users of their services. Their policies apply globally, without reference to a specific legal framework, even if some OSPs use the term 'illegal content'. To be considered illegal, content must indeed violate a specific law. But almost none of the OSPs in the sample refers to a specific legal framework from a country or a supranational entity to specify what content is illegal.

This does, however, not mean that the legal framework is not taken into consideration for other purposes, at least by some OSPs. First, different versions of terms and conditions exist for the same platforms, whose application depends on the user's residence. It should be noted, though, that the distinctions do not relate to the criteria determining whether content is permissible or not; these criteria are identical for all users. The T&C only differ to account for specific aspects of locally applicable regulations, such as consumer law. Second, some OSPs refer to specific legal framework concerning the consequences applicable to (alleged) illegal content. For instance, certain OSPs (e.g., Reddit) refer to the Germany's Network Enforcement Act (NetzDG law). More broadly, OSPs also refer to the national framework when referring to their cooperation with LEAs, regarding orders to remove illegal content, on the understanding that such an order, if applied, will produce its effects on the national territory of the issuing authority, i.e., where the content is deemed illegal, while the content continues to be accessible elsewhere. Third, some OSPs recall their commitments to fundamental rights (especially freedom of expression), with a view to avoid removing content if this operation would limit freedom of expression or to allow content which would otherwise go against their standards, because it is newsworthy and in the public's interest, or because the speech is delivered in a specific context. Such commitments to fundamental rights can lead to a narrow interpretation of government requests or to a different application of the community guidelines to politicians or government officials. Fourth, several OSPs refer to the norms of a country or region (without specifying which one) when they mention exceptions in the guidelines. Consequently, while the general rule applies globally, exceptions to it may differ across regions. For instance, TikTok states that changes in the law (again, without specifying which one) can lead to changes in the policies. Snapchat is even more explicit, since it requires its users to comply with local, state, national and international laws, rules and regulations (so does, e.g., Facebook) and explains it takes action against any activity that undermines public safety, U.S. laws or the laws of the country where the user is located.

Permissibility criteria related to NCII

With respect to NCII, seven permissibility criteria can be distilled from the data sample. In most cases, it is the combination of those different criteria that will lead the OSP to decide whether to remove the content.

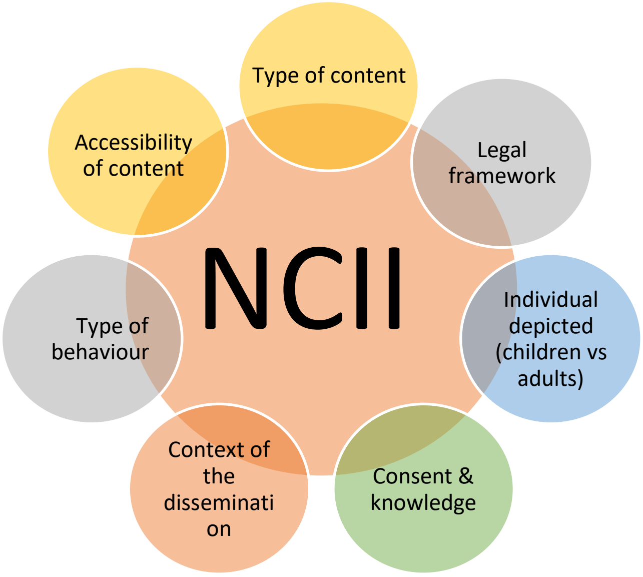


Figure 23. Permissibility criteria related to NCI

1) Type of content

The type of content inevitably plays a role in determining whether it is permissible. Interestingly, most platforms do not refer to 'intimate images' as such, with some exceptions (e.g., Twitter refers to 'intimate photos or videos').

Some platforms refer to the sexual nature of the content with sometimes a zero-tolerance policy to prevent any non-consensual or underage sharing of content. It can refer to sexual activity, sexual pose or to pornography, understood in a sense that includes depictions of intimate parts of the body clothed or not engaged in a sexual activity. In some cases, content that depicts simulated or implied sexual acts (LinkedIn) are regarded as content of sexual nature. This can also be the case for other kinds of content than images and videos, such as audio content erotic literature.

Most OSPs in the sample also use the terms of 'nudity' or 'nude images' to refer to content which does not depict sexual acts. These terms relate to nude pictures, generally depicting commonly sexualised body parts (such as genitals, breasts, groin, buttocks, thighs), including pictures taken via creepshots (i.e., a picture taken surreptitiously and without consent, usually (but not always) of a woman, and usually focusing on sexualised areas of the body) or upskirting (i.e., a picture taken up a person's dress or skirt without permission) or close-ups of clothed body parts and people in see-through clothing.

In the same way, several platforms take into consideration (at least partially) created or manipulated content (deepnudes), drawings, sexually explicit language, links to non-permissible content, thumbnails, banners, avatars and emojis.

In some cases, it was observed that the private or commercial nature of the content plays a role in the appreciation of its permissibility: non-commercial imagery or imagery produced in a private setting can be impermissible, in contrast with commercial imagery. Nevertheless, others also prohibit commercial imagery.

2) Conformity with the law

As indicated above under the general findings, OSPs sometimes refer to legal framework as a permissibility criterion. This is specifically recalled by Pornhub, which requires the creator of the content to warrant that it does not contravene to any applicable laws and does not subject Pornhub to any claims, demands, lawsuits, regulatory actions or any actual, potential or risk of liability or any threats thereof.

3) Minors or adults

Most selected platforms distinguish between adults and children or minors depicted in the image. Many of them have a zero-tolerance policy when it comes to content depicting minors, regardless of whether it is shared in a consensual manner or not, even if it represents teenagers engaged in a sexual conduct in a consensual way. The terminology used is far from uniform: some platforms refer to sexual exploitation of children, to child sexual abuse, or to behaviours that put young people at risk of exploitation or psychological, physical, or developmental harm. Neither do platforms necessarily define what is meant by a child, although most agree to include an individual who appears to be a child. Some refer to people under the age of 18, or even to people older than 18 if the individual represented is in a location where 18 is not the age of majority.

4) Consent and knowledge of the depicted person

Consent is also a central criterion when it comes to assessing permissibility of intimate/sexual materials, especially for adult content. If the individual depicted did not consent, the sharing of this content is not permissible. Some platforms require consent to be given for acts, recording, dissemination, and manipulation of the content. Pornhub defines consent by referring to 'the express, voluntary, and non-coerced agreement or willingness to engage in a specific sexual activity and, where applicable, to produce or disseminate content for a particular audience. It is the power individuals have over their bodies, images, and the content they generate or of which they are a part'. It is not enough to assume to have a valid consent: 'consent must be determinable by the reasonable observer from the material itself, through clear verbal or visual cues'. Most other platforms prefer to give examples of situations where consent is deemed to be lacking, such as in a revenge context (which depends on the caption, comments and/or a page's title), on the ground of information received from independent sources like LEA, if the person depicted flags the content, or depending of the context in which the content was created (hidden camera, creepshots or upskirting, deepnudes, images or videos created in an intimated setting, or in the case of doxing, persons appearing drugged, incapacitated, intoxicated or asleep).

In addition to consent, one platform (Discord) refers to the knowledge of the individual depicted stating that sexual content is allowed but not without his/her knowledge. This brings to mind the Belgian legal framework which punishes NCII without the consent *or* the knowledge of the person depicted.

5) Type of behaviour

Most platforms prohibit the sharing of content containing intimate images. Some platforms also consider as impermissible prior behaviours like threatening, expressing an intent to share, offering, or asking for non-consensual intimate imagery, or offering a financial reward or bounty in exchange for intimate videos or pictures. When it comes to content depicting minors, impermissible behaviour is even defined more broadly and includes the attempt to obtain sexually explicit content, viewing and linking to impermissible content.

6) Context of the dissemination

The context of the dissemination will sometimes be taken into consideration as a permissibility criterion. Some OSPs indeed consider sexual imagery or nudity as permissible content if it is posted for educational purposes, for humorous or satirical purposes, for protest, to raise awareness about a cause (for instance, child nudity in the context of famine, genocide, war crimes or crimes against humanity), for medical reasons, or for health reasons, at least if such intent is clear. Other OSPs add content related to cultural events, to historical events, at least if there is a public interest to view the content (Snapchat). Most of OSPs have also created an exception for art (e.g., Facebook, which permits photographs, painting, sculptures, and other art depicted nude figures).

The user's purpose to trigger sexual gratification is also a criterion taken into consideration. The intent of sexualising the body or portraying an individual in a sexual manner, or the salacious manner in which a person is depicted, can also be taken into consideration, even if, for instance, it concerns pictures of clothed children not engaged in overtly sexual acts. On the contrary, the non-sexualised way in which a minor or a person is depicted can lead to the content being deemed as permissible.

In addition to the context in which the publication is published by a user, the context faced by the user likely to see the content can also be taken into account. This refers to the user's age, location, and preferences, also depending on the culturally accepted practices (such as indigenous populations). Moreover, nudity can also be defined by reference to the cultural norms: TikTok, e.g., states that nudity is showing intimate body part that prevailing cultural norms indicate should be fully covered.

Depending on the targeted audience, the permissibility assessment may also vary, especially when content targets young minors and families (YouTube, which prohibits violence, obscenity, or family-friendly cartoons engaging in inappropriate acts).

7) Accessibility of the content

In certain situations, and generally depending on the type of content, certain platforms authorise the publication of intimate images with restrictions regarding its accessibility. For instance, some OSPs restrict the display or prevent broad dissemination, issue warning labels. Other apply adults-only restrictions.

Permissibility criteria related to OHS

With respect to OHS, it is also possible to identify seven permissibility criteria. While there is some common ground with NCII, the criteria applicable to OHS are not fully identical.

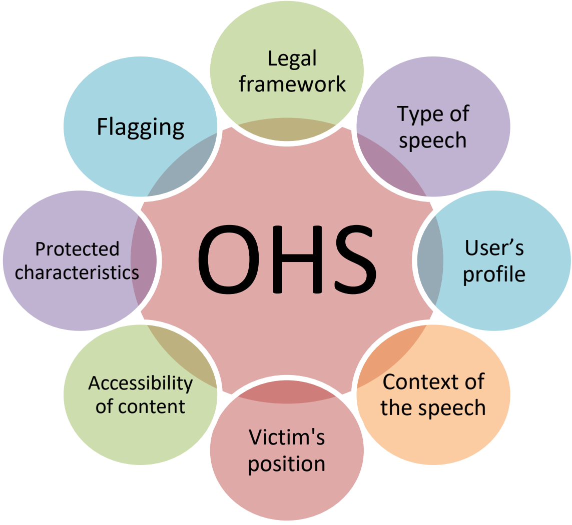


Figure 24. Permissibility criteria related to OHS

1) Type of content

There is a wide variety among platforms in the wording, definition, and scope of what is considered non-permitted speech. All platforms target hate speech, but a wide variety of (sub)labels to refine the impermissible content or behaviour.

Furthermore, the OHS policy of certain platforms does not only encompass specific content but also an ideology or a belief, such as Nazism, white supremacy, white nationalism, or white separatism. As a result, all content related to such an ideology or belief is impermissible, regardless of the exact content of the speech. It is worth mentioning that all platforms (with the exception of BeReal, Telegram and Reddit) explicitly prohibit negationism in various forms, including denial of the Holocaust or other genocides, the misappropriation of holocaust symbology, or the denial of other events like slavery in the U.S.

Except for negationism, the above-mentioned types of speech are not in themselves unacceptable. Most platforms consider speech to be impermissible if it meets both the targeted type of speech and (at least) one of the protected characteristics, explained in the following subsection.

2) Protected characteristics

Of the selected platforms, all have a specific policy on hate speech, with the exception of BeReal and Telegram. Those that have developed a specific policy on hate speech, rely on specific protected characteristics. Interestingly, these characteristics are not necessarily the same for all OSPs. However, all OSPs protect the most widespread characteristics, by prohibiting speech based on racism and xenophobia, sexual orientation, gender, disability, and religion (with the exception of Reddit regarding this last characteristic).

Most platforms also target other characteristics, for example pregnancy, size, source of income, caste, serious decease, age, belief, and certain statuses like kin, socioeconomic, veteran, or housing status, the status of victim of domestic and sexual violence or stalking or of a major violent event. In some cases, characteristics other than those expressly indicated by the platform can also be taken into consideration. In such cases, rather than considering such content or behaviour as hate speech, some platforms regard it as harassment, at the very least, when it targets an individual.

3) User's profile

Certain platforms do not only target problematic content, but also prohibit certain persons or entities from using the platform, such as hateful entities, organisations or groups, dangerous and hateful individuals. In some cases, a repetition of the hateful conduct (Facebook, Instagram), online or offline, is required, and a thorough background check is provided before banning the account (TikTok). Moreover, Twitter excludes from the prohibited entities state or governments entities, as well as representatives elected to public office, even if they can be considered as hateful persons or entities. That said, the platform also states that there is no place for hateful entities, e.g., entities that systematically and intentionally promote, support and/or advocate for hateful conduct. By contrast, other platforms (like YouTube) state that they apply their rules equally, regardless of the user's background, political viewpoint, position, or affiliation.

4) Context of the speech

The context of the speech is very often taken into consideration before regarding it as impermissible. Some platforms, such as Facebook and Instagram, accept satirical content or content posted to mock or criticise someone or something. In contrast, Reddit, for instance, prohibits community groups (or 'subreddits') that are dedicated to mock other people.

A couple of OSPs apply a 'public interest exception' for content that would otherwise violate the standards, but that is in the public interest, as it relates to a topic that can inform, inspire, or educate the community, or because it contributes to the understanding or discussion of a matter of public concern. Nonetheless, Twitter adds strict conditions to authorise the publication of such posts: they can only be posted on a verified account of a current or potential member of a local, state, national or supranational government or legislative body, with more than 100.000 followers. Moreover, the content will be hidden behind a warning screen providing context. YouTube applies a similar exception for 'EDSA content' (i.e., educational, documentary, scientific or artistic content), which refers to videos that might otherwise violate the terms of use but that are permissible on the ground of a 'compelling reason', appearing in the image or audio of the video itself and in other items like the video's title and description. Content can also be permissible when the user's explicitly mentioned reason is to condemn hate speech or slurs or raise awareness among other users (Facebook & Instagram). Finally, a few platforms expressly state that they account for the cultural, linguistic, or regional context in which the content is published.

5) Victim's position

The victim's position too can play a role in determining whether speech is impermissible. While certain platforms exclude OHS against concepts or institutions, other platforms are, or at least appear to be, more tolerant when the target of the speech is a public figure, a person in a position that receives a lot of public attention and having ways to counter negative speech, at least if the speech expresses a critique that is in the public interest. In some T&C, violent threats against law enforcement officers are also considered impermissible content. When an individual is the target of a comparison to animals or insects culturally perceived as intellectually or physically inferior (e.g., monkeys or cows) or of terms related to female gendered cursing, such content is not categorised as OHS but as harassment. As such, it remains impermissible content.

6) Accessibility of the content

It is important to mention that one of the platforms in the sample makes a (general) clear distinction in terms of content permissibility based on the audience that is likely to be reached by the content. The assessment of permissibility will be stricter when the content is eligible for 'FYF' (i.e., For Your Feed, a unique TikTok feature that uses a personalised recommendation system to allow each community member the ability to discover a breadth of content, creators, and topics). In contrast, content distributed to a smaller audience is subject to a less stringent assessment and may therefore be considered permissible.

7) Flagging by users and trusted flaggers?

Interestingly, the analysis shows that the role of flagging seems to be marginal. Only a few platforms expressly indicate that the flagging by a user or trusted flagger can play a role in the assessment of permissibility. For instance, Instagram regards racial slurs as permissible content, but it can be considered inappropriate, disrespectful, or offensive by a user. If the latter flags the content to the platform, the content will, assumably, be regarded as impermissible more easily.

ii. Survey among moderators

Whereas the OSPs define the general T&C and policies, their teams of (individual) moderators are confronted daily with the assessment of online content as permissible or impermissible, whether it is flagged by an AI tool or by a user, and they decide on the action to be taken. Therefore, moderators are key enforcers of the self-regulatory framework analysed above. Their actions have a direct impact on online content (Suzor, 2017). For this reason, the team designed and performed a survey, including a questionnaire and a vignette-study, among moderators.

Literature review

In a first step, it appeared crucial to conduct a review of the existing literature on content moderation and the role of moderators. This literature review provided a solid understanding of the subject matter. The key elements that emerge from the literature review are as follows.

According to literature, content moderators have diverse educational backgrounds and rarely come with a legal background (De Gregorio, 2020; Dias, 2020). While most have a college degree, they do not come from the typical 'hard sciences' such as engineering, computer science or medicine, but rather from 'a variety of liberal arts and humanities fields' (Roberts, 2021).

Training programmes depend on the tasks and content the moderators will be given and are based on extensive, detailed documents. Some researchers refer to a 'bible' of 10,000 words, composed of 24 different categories, that moderators are instructed to follow (Wilson & Land, 2021). Whistle-blowers mention a combination of several documents, including a policy document of 30,000 words split up into various webpages, a document with 'known questions', another one with 'operational guidelines' and a 'workplace' containing 'one-off exceptions' (Hartwig & Heckenlively, 2021). This documentation is frequently updated, which requires moderators to refresh their knowledge regularly and undergo tests to demonstrate their understanding of the internal policies (Gerrard, 2020). Moderators' training would however present several shortcomings (Bellanova & Goede, 2020; Gill, 2021), such as a lack of feedback on emergency and safety protocols (Windler et al., 2019).

Literature also highlights several factors that can lead to a bad decision, including working conditions, language and cultural considerations, and the existence of grey zones in content moderation. Together, these factors contribute to the complexity of the moderation process and underscore the challenges faced by those responsible for maintaining online platforms' content standards. First of all, content moderators reportedly face significant challenges in their working conditions. One of the major issues is the limited time they are given to make decisions (Aswad, 2018; De Gregorio, 2020; Gongane et al., 2022; Siapera, 2022). Studies indicate that the time allocated to a moderator to assess content is often too short, providing inadequate room for thoughtful consideration. In fact, certain studies mention a time frame as brief as one minute for decision-making (Wilson & Land, 2021) or even 10 seconds (Gillespie, 2018). Second, in the process of determining whether content is harmful or not, moderators must possess a deep understanding of the nuances of various vocabulary words (Dias, 2020). This complete comprehension is challenging, even for moderators who are moderating in their native language. Cultural and linguistic intricacies can make it difficult to accurately interpret the intent and meaning behind certain phrases and words. This is especially true when dealing with idiomatic expressions, regional slang, or cultural references that may not be familiar to the moderator. As a result, even native speakers can struggle to accurately assess content, leading to potential misunderstandings or misjudgments (Wilson & Land, 2021). Third, moderators are to deal with so called 'grey areas', i.e., situations or content not clearly defined by moderation rules or policies. These are cases where there might be subjective or ambiguous interpretation on whether the content violates the rules or not. Due to the complexity and subjectivity of these cases, they can often pose challenges for moderators when deciding whether they should be removed or allowed (Paasch & Strippel, 2021; Pöyhtäri, 2014; Conseil de l'Europe, 2021). As Kaye (2019) rightly states: 'Drawing the lines is hard.' Since certain content is not governed by clear rules, this leaves room for a certain degree of subjectivity (even if the OSPs' participating in the roundtable argued otherwise, see above, b.). Literature raises questions about the neutrality of moderators (Conseil de l'Europe, 2021), depending on their personal experiences and sensitivities (Siapera, 2022). According to a 2017 study, for instance, there seem to be differences between male and female moderators regarding the labelling of comments as 'toxic' (Binns et al., 2017). Biases in OSPs' policies have been criticised as well (Hartwig & Heckenlively, 2021), just like the 'American or Western cultural lens' through which many OSPs (especially U.S.-based OSPs) look at online content (Roberts, 2021); some even refer to 'digital colonialism' (Kaye, 2019).

Notwithstanding the existence of a considerable body of literature, our understanding of the dynamics that shape decision-making processes in the field of content moderation remains, all-in-all, rather limited. It indeed continues to be difficult to grasp the factors that influence decision-making in content moderation. This is due to the high degree of confidentiality and the opacity of internal rules. Moreover, while prevailing research tends to emphasise the negative aspects of moderation, gaining a better understanding of moderators' aspirations and the factors contributing to their commitment to this profession would be highly enlightening. Furthermore, it is important to note that existing research is based on (very) small samples of interviews or testimonies, and primarily focuses on a small number of large OSPs or platforms (especially Facebook, YouTube, Twitter). Indeed, most literature only focuses on a few major players in this field and thus has a limited scope, which further constrains our understanding of content moderation processes in a broader context. In sum, there is a clear and urgent need for more in-depth qualitative research in this field.

Target population of the survey and recruitment

To match the selection of OSPs made for the analysis of the self-regulatory framework, the team decided to focus on online content moderators operating in the EU market (instead of the Belgian market, as initially envisaged in the project proposal). Since the overall objective of this survey is to better understand the application of OSPs' self-regulatory framework, the survey was limited to professional moderators; non-professional (community and other) moderators were deliberately excluded as they do not necessarily apply (or not only) the platform's policies.

Considering the secrecy or 'opacity' that reigns in the online content moderation realm (Gillespie, 2018; Kaye, 2019; Roberts, 2021), the team decided to implement a double strategy: on the one hand, we reached out to individual moderators, while on the other hand, we asked the OSPs that had participated in the stakeholder meeting to participate in the survey by implicating, ideally, their own moderator teams or, alternatively, their compliance team. Eventually, this alternative option was preferred by the OSPs. The survey's sample thus consists of two subsamples: individual moderators (A) and OSPs (B). This two-fold approach allowed us to gather insights from the perspective of both moderators and OSPs.

Reaching individual moderators (i.e., subsample A) proved to be extremely challenging and time-consuming, despite the implementation of various strategies. To recruit participants, several methods were applied: (i) targeted advertisements on Facebook; (ii) public posts published by all team members on LinkedIn and/or Twitter; (iii) more than 100 private messages sent to individual moderators using professional LinkedIn accounts; (iv) snowball sampling technique; and (v) emails and LinkedIn messages sent to several recruitment companies that are known for hiring online content moderators, as well as associations and unions of moderators (or more broadly, the tech industry). Despite these strategies and several reminders, the team faced numerous rejections, especially from individuals invoking the confidentiality clause in their work contract, but above all, an overwhelming rate of non-responses.

As to subsample (B), the OSPs that had participated in the roundtable stakeholder meeting or in one-on-one meetings were recontacted via email. Other OSPs, which had been contacted for the roundtable and showed interest in the @ntidote research project but eventually could not attend the meeting, were recontacted via email as well. Some did not reply, even after one or two reminders; others refused to cooperate, indicating all information on content moderation can be easily found on their website. None of the contacted OSPs agreed to sharing the survey with their moderators. A few accepted to submit the survey to their compliance team. Two OSPs eventually lived up to this commitment.

Design of the survey

The survey seeks to understand how content moderation is organised and functioning in practice (including training of moderators) and how moderators/OPSs delineate (im)permissible online behaviour. Furthermore, it aims to investigate the interaction between automated tools and human moderation in detecting and dealing with (possible) impermissible content, and to learn about the possible reactions to impermissible content in practice. To this end, the team developed a questionnaire and a vignette-study. Whereas the questionnaire specifically targeted individual moderators (i.e., subsample A), the vignette-study was addressed to both subsamples.

The questionnaire was developed based on the literature review. Since the target population consisted of professional online content moderators operating on the EU market, the questionnaire was drafted in English (see annexes 18 and 19). The main aim of the questionnaire was to delve deeper into the realm of moderation and to test the findings resulting from the literature study. To this end, the questionnaire includes a combination of open-ended and closed questions covering specific facets of the respondents' experiences, including moderator profiles, training protocols, challenges faced, and the content moderation process as experienced by moderators themselves.

The vignette-study encompasses a series of scenarios on OHS and NCII, each accompanied by closed questions. The scenarios were selected from the ones used for the vignette-study in WP3, with the aim of comparing how moderators and digital natives assess the permissible/impermissible nature of online content, and more in particular OHS and NCII. In other words, the objective is to investigate the black box of content moderation to better understand the moderation decision-making process, the role of AI tools in this process and the factors that influence the permissibility assessment made by the moderators (e.g., skin colour, sexual orientation). Compared to existing literature, showing that the moderation decision-making process is 'opaque' (Gillespie, 2018) and may be experienced as inconsistent or strange (Roberts, 2021), the vignette-study is highly innovative as it is, to our knowledge, the first time that such study has been undertaken with respect to OHS and NCII. The design of the vignette-study was a deliberate attempt at creating a virtual environment simulating the complexities of real-time moderation, notwithstanding the constraints regarding the selection of scenarios to enable a comparison with the results from WP3. The scenarios encompass diverse situations, prompting participants to navigate the nuanced terrain of content evaluation, including decisions on content removal and other suitable sanctions.

Before launching the survey, the team performed a pre-test to check the feasibility of the survey (especially in terms of length/time) and to make sure that all questions and scenarios were clear.

In developing the survey, the team committed to strong data protection and confidentiality. All participants were guaranteed anonymity and confidentiality. No personal data, such as names, surnames, or IP addresses, were collected during the survey. Furthermore, we have taken stringent measures to ensure that the dissemination of results will never include individual data, thereby eliminating any possibility of identifying participants through the aggregation of responses. This approach not only upholds the principles of ethical research but also safeguards the anonymity and confidentiality of those who participated in the study.

Sample

Despite all the efforts to reach both subpopulations, the eventual sample is quite small. Therefore, one needs to be careful with drawing general conclusions. That said, the sample is quite diverse, as the presentation below shows. In this respect too, the @ntidote survey distinguishes itself from existing studies and literature which, as indicated, mainly focus on a few big OPSs. Below, the numbers between round brackets represent the number of replies or instances.

Subsample A comprises 13 respondents: 6 men, 5 women, and one individual identifying as non-binary/genderqueer. Two of them were former moderators. Their birth years span from the 1970s to the 2000s, with six born in the 1990s and four in the 1980s. This indicates that, for many, content moderation is not a first or 'entry-market' job, contrary to what is suggested in some literature (Roberts, 2021).

In terms of education, eleven participants indicated having pursued higher education, including professional bachelor's, academic bachelor's, or master's degrees. One participant only possessed a secondary school diploma. The array of academic disciplines is notably diverse, encompassing fields such as Medicine, Business Administration, Political Science (2), Economics, Anthropology, French Language Teaching, Civil Engineering, Advertising, and Computer Science Engineering. It is notable that these findings align with existing literature, which also emphasises the diverse educational backgrounds of moderators (Dias, 2020). Interestingly, the sample also includes moderators with a background in 'hard sciences' (Medicine, Engineering), contrary to what the literature study revealed. The sample also seems to exhibit a higher level of education compared to what is observed in existing literature. This divergence could be attributed to our recruitment strategy via LinkedIn, which might have attracted individuals with higher levels of education. Finally, it is also interesting to note that one person did not respond to these socio-demographic questions.

Furthermore, our respondents were moderators for (at least) 6 different OSPs/platforms, mainly but not only U.S.-based; 4 of them preferred not to answer. They had been recruited by various companies.

The language of moderation varies considerably. From a list entailing all official languages of the EU, English was chosen by nearly all respondents (12), followed by Italian (3), Spanish (3), Dutch (1), Greek (1), Hungarian (1), Polish (1), and Portuguese (1). Interestingly, 5 moderators ticked the category 'Other(s)'. One respondent preferred not to answer. In addition to being diverse, these answers show that English is the 'lingua franca' and that all participants moderate in more than one language, which connects to the earlier identified language issues.

While the respondents were not asked explicitly to describe the type of content they usually deal with, they were invited to give an estimate of the percentage of cases that relate to OHS and NCII. We can observe that our respondents, on average, encounter approximately 25-49% of content related to OHS and 5-9% of content related to NCII.

Finally, subsample B consists of two major OSPs.

Analysis of the survey results

Despite the small sample, the data collected through this survey is particularly rich. This report only presents the data and results that are most relevant for the research objectives.

1) Training of moderators: a rather positive image

According to the respondents, the training duration ranges from 2 weeks to 1 month. It is provided by a recruitment company in 12 cases, while 5 moderators indicated they had (also) been trained by the online platform (or the OSP). The training programmes consisted of a combination of on-site training (10), online training (7), and self-study of a course (e.g., manual, slides) (3).

When asked about the adequacy of the training, 6 participants answered that '[t]he training was sufficient', seven responded that '[t]he training was sufficient, but with some gaps'. Interestingly, none of the respondents indicated that the training provided was insufficient. The open-ended responses highlight the following four elements: the ease of the work and the comprehensiveness of the training, but also the lack of competent trainers and difficulties related to borderline cases.

2) Content moderation: a diverse reality

How does content end up on the moderator's desk? Moderators obtain content through various channels: automated tools (10), non-professional content moderators (3), users (10), law enforcement authorities (4), and colleagues (1). One person preferred not to answer. The volume of moderated content per day seems to vary tremendously, between 40 so-called 'tickets' and 2.000 pieces of content, depending on the type of material (text, image, video). Our respondents process an average of 850 pieces of content per day. This high workload confirms the findings based on the literature study, concerning the importance of speedy responses to meet OPSs commitments to its users (Gillespie, 2018). More specifically, respondents indicate having 10 seconds to 30 minutes per case (cf. De Gregorio, 2020; Wilson & Land, 2021), but these figures seem to vary significantly: 'It depends on a lot of factors, I couldn't give a straight answer'; 'It depends on the project, but we can work just 2 hours per day on content that is considered "high severity" (child porn, suicide content, etc.)'; 'It depends on the kind of content. Easy policies don't take longer than a few seconds. Complex policies take up to 20 minutes or more'; or simply, 'It depends on the content.'

The study also focuses on what happens to the content pending the moderation decision, with proposed answers in the form of a list of questions based on existing literature. The treatment appears to vary substantially: the content remains online (2), the content is temporarily removed (3), the content is tagged or labelled in a specific way to inform users (4), or it depends on the type of content (3). Two participants preferred not to answer. Furthermore, 3 moderators indicated that they do not know what happens to the content in the meantime.

Regarding the feedback moderators receive, it is worth noting that this feedback would primarily focus on quality (12) or, to a minor extent, both quality and quantity (1). These results are rather surprising, considering the challenges identified based on existing literature (not enough moderators, time pressure) and the high daily volume of content to moderate. Moreover, on the one hand, these answers contradict the literature that suggests moderators are required to meet a certain quantity of content moderation to maintain their employment (Arsht & Etcovitch, 2018). On the other hand, they conflict with some of the answers given by the participants when asked about the challenges their job raises, as the next subsection shows.

3) Challenges and tensions experienced by moderators

The moderators were also questioned about the challenges or difficulties they encounter while exercising their profession. The proposed answers in the form were a list of questions based on existing literature (Arsht & Etcovitch, 2018; Aswad, 2018; Wilson & Land, 2021). The responses were as follows: I do not (or not always) have sufficient time to make solid decisions (3); there are not enough moderators to deal with all the content to moderate (4); I am not sufficiently trained for the job (0); the platform's policy rules are not easy to apply in practice (4); I have difficulties understanding the language of the content to moderate (3); I have difficulties understanding the culture of the users whose content I moderate (1); I do not receive the necessary psychological support (1). Finally, three respondents said they experienced no difficulty at all.

When asked to elaborate on their answer, the respondents highlighted the following positive aspects:

- ‘Gathering human activities under appropriate labels is the nature of our work, and for that, we must know the culture and language’;
- ‘The job is not difficult at all; just keep updated on the new policy and knowledge of the market you are working for’;
- ‘We have absolutely everything to perform well in this role, and there are no expectations in terms of time’.

They also pointed out some negative aspects:

- ‘Management gives specific seconds to answer to a video which is never enough to make a right decision. Policies change daily’;
- ‘Unrealistic quality and quantity standards, on purpose so the company can rotate employees frequently by not renewing contracts’;
- ‘I believe this work can lead to several mental health issues, the support on-site is useless’;
- ‘We are often asked to moderate content in other languages using a translator software’;
- ‘We moderate several areas but are only a few moderators’;
- ‘Sometimes is something about visibility or a word that even natives don't know what it means that makes the job hard’.

From the additional explanations given by the participants, it also emerges that frequent policy changes can pose a challenge (cf. Wilson & Land, 2021). Regarding the question about the policy update frequency, the responses are as follows: ‘reviewed weekly’; ‘every day’; ‘2 weeks’; ‘every 2 weeks’; ‘I'm not sure I can disclose that information’; ‘very frequently’; ‘every week the client makes an interview with our Q&A team. I can't answer’; ‘1-3 months’; ‘every week’; and ‘it depends’.

Next, when we inquired whether they have experienced tensions between the decisions they must make under the platform's policy rules and the decisions that appear fair to them regarding NCII and OHS, five respondents answered ‘no’, two stated ‘quite often’, five mentioned ‘sometimes’, and one preferred not to answer. When illustrating such tensions, the moderators explained that the tensions could be related to specific words or ideologies (e.g., ‘For hate speech, I disagree with the fact that if you talk bad about LGBTQ as an organisation, it is considered an attack toward the community’).

Despite the existence of tensions, the solution seems unanimous: all moderators believe it is necessary to follow the company's policies (e.g., ‘I follow the company's policies because they have told us that our opinion doesn't matter’). This confirms earlier research showing that company rules prevail even if they clash with the moderator’s own values (Roberts, 2021; Kaye, 2019), and the outcome of the roundtable stakeholder meeting (see above, b.). However, in some cases, the decision-making process may involve team discussions (e.g., ‘We talk with the law enforcement team’). This too confirms earlier research (Seering, 2020); more in general, team meetings to discuss policy frequently take place, even if the input from individual moderators does not seem to have a big impact on the policy rules (Roberts, 2021).

4) Confidentiality: to disclose or not to disclose?

The survey suggests a significant level of confidentiality in the moderation decision-making process, thereby clearly confirming existing literature. From the survey recruitment process, it was already obvious that conducting research on online content moderation is not an easy task. Not because the results of the survey would be revealing major flaws – at least, this cannot be said for the above-mentioned findings. Neither because the population is, in theory, difficult to access. But primarily because there exists a contractual barrier between moderators and researchers: non-disclosure agreements or confidentiality clauses.

With only 13 respondents, several repeatedly stated 'I prefer not to answer'. For instance, one of the respondents answered about 20 questions (out of 100 in total) and 9 of the replies were 'I prefer not to answer'. In total, 73 'I prefer not to answer' responses were given, even to very basic and/or seemingly less sensitive questions such as: 'During the moderation process, what generally happens with the content pending the decision?'; 'Would Arthur's comment be flagged by the automated tools deployed by the online platform(s) you moderate for?'; 'Would there be any other reactions possible, other than or in addition to removing Arthur's comment, on the basis of the platform's policy rules?'; 'Do you, in your role as a content moderator, sometimes report content to law enforcement authorities?'; 'Do you have the feeling that your work as a content moderator is appreciated?'.

Furthermore, several respondents explicitly stressed the confidential nature of the data. Some did so when answering the final question ('Is there anything else you would like to add?'), others signalled this elsewhere. For instance:

- 'The topics covered within this survey, although surely being of great public interest, are considered confidential information over the company's business practice. If I asked to my supervisor to participate in this survey, it should've been reviewed by the legal office first and probably they would've told me not to participate at all. Also, my request could have caused prejudice against my ability of abiding to the confidentiality norms. For sure, if identified as a participant I will receive a formal investigation and probably my employment would be terminated';
- 'Although I am forbidden to share company information, I wanted to fill out your form because I think academic studies are more important than company sensitivities';
- 'Not allowed to discuss, prefer not to say, no'.

Even though it is widely known that moderators sign a non-disclosure agreement that prohibits them from revealing company information, one may question the reasons behind this extreme confidentiality (Drootin, 2021). Existing literature provides several explanations (Roberts, 2021): (i) companies want to protect their moderation policies and practices as proprietary information, on the one hand, to avoid users from attempting to game the rules and, on the other hand, to give themselves a competitive edge; (ii) this way, OSPs try 'to escape scrutiny and public review of these policies'; and (iii) content moderation is 'an unpleasant necessity', therefore companies prefer to keep this process as invisible as possible. Moreover, we also know that company policies are not neutral, but are drafted through a certain 'cultural lens' - often American or Western, even if that is changing with the entry into the market of OSPs like Telegram and TikTok (Roberts, 2021) - and that they may fluctuate depending on the social climate or due to specific events (for some striking examples, see Kaye, 2019).

Still, considering the survey was anonymous and confidential treatment of the results explicitly guaranteed, one wonders what drives the respondents to refrain from answering certain questions, even though most responses were positive for the OSP, such as comprehensive training, low tensions, or limited difficulties. Why do so many moderators (or former moderators) adhere so strictly to non-disclosure agreements? Several theories can attempt to address this question:

Rational Choice Theory (Cornish & Clarke, 1986)
According to this theory, individuals make decisions based on a cost-benefit analysis. Moderators may consider that the potential benefits of disclosing information do not outweigh the costs, such as job loss, legal action, or damaged reputation. Rather than participating in a scientific survey which creates few direct benefits for the moderators, they prefer to adhere strictly to the non-disclosure agreement. Furthermore, the notion of rationality seems highly relevant when discussing the work of moderators. As the survey results demonstrate, some respondents describe their work as 'easy' because it simply involves adhering to the policy. In case of tensions between their personal opinions and the policy rules, the latter clearly prevail. There is no room for subjectivity, emotions, or personal ideologies (cf. the roundtable discussion with industry). Rationally, they must execute the choices made by the company, whether reflected in the policy rules, their interpretation by moderators' superiors, or the solution worked out with the law enforcement team. They are told not to think too much (Kaye, 2019). Finally, one of the factors influencing this rational choice could be social control.

Social Control Theory (Hirshi, 1969)
This theory suggests that individuals are influenced by social control mechanisms around them. Moderators may face tight control from their employers or other powerful actors, which discourages them from disclosing confidential information. It can be assumed that the control experienced by different moderators in their daily work (feedback on quality, speed of moderation, double-checking, etc.) can amplify the sense of social control, regulation, and the fear of punishment, thus potentially leading to a loss of benefits.

Social Learning Theory (Bandura, 1976)
This theory posits that individuals model their behaviour by observing the rewards and punishments of others. As early as 1993, Cusson discussed the structuring nature of social control. If violations of non-disclosure agreement have previously resulted in negative consequences (warning, termination, legal actions, reputation issues, etc.) (for an example of such punishment, see Hartwig & Heckenlively, 2021), it can deter or discourage other moderators from following the same path. Observation of the negative consequences faced by colleagues may indeed serve as a deterrent. Respecting non-disclosure agreements becomes a form of 'modelling' that becomes the norm within the company. It establishes a desired status, such as earning respect from colleagues or superiors, potential career advancements, and other benefits. This, in turn, influences moderators to adhere to the agreement.

These are, of course, only plausible theories. To assess what exactly explains moderators' strong adherence to non-disclosure agreements, further research would be necessary (e.g., a survey followed by in-depth individual interviews).

5) Factors guiding the assessment of (im)permissible online content by moderators

Based on the second part of the survey (i.e., vignette-study), some interesting conclusions can be drawn on the assessment of (im)permissible online content by moderators, even if one should remain cautious with general conclusions considering the small sample.

It is noteworthy that several factors do *not* seem to exert a significant influence on this permissibility assessment, even if the team's research assumption was the opposite:

- **Flag origin:** regardless of whether the content was flagged by an automated tool or by the victim or another user, the responses exhibit striking similarities, sometimes even being identical across all scenarios. Nevertheless, scenarios involving NCII appear to be slightly more prone to detection by automated tools and are more likely to be removed (before a user can even flag the content).
- **Content removal**: regardless of the scenario, the responses highlight a significant inclination towards considering content removal when it is flagged by an automated tool.
- **Importance of automated tools**: the automated tool would play a crucial role for all the scenarios, even slightly more so for NCII than OHS. This confirms existing research on the importance of AI in content moderation (Castets-Renard, 2020), even if AI has more difficulty with text and context (Kaye, 2019), and the outcome of both the roundtable stakeholder meeting (see above, b.) and the analysis of technical tools used by OSPs (see below, d.).
- **Assessment of the (im)permissible nature of OHS**:
  - ● With respect to the influence of gender, only two moderators indicated a potential difference in their responses for scenario 2b. One respondent differentiated the grounds for removal (2a: hate speech, 2b: cyberbullying). In subsample B (**OSPs**), while the responses of both OSPs are notably the same for all scenarios, this does hold true for scenario 2. Whereas one OSP indicated to be unsure about the responses, the other OSP replied that the content would be flagged and removed if it targets a girl; in contrast, if it targets a boy, it would only be removed upon the request of the victim.
  - ● With respect to the impact of the skin colour, only one respondent highlighted that the comment might be removed solely if the young white man directed hateful remarks towards the young black man.
- Assessment of the (im)permissible nature of **NCII**: in regards to the victim's **sexual orientation**, just one moderator gave different answers for scenario 4b, but at the same time this respondent stated that the automated tool should flag the content and that it should be removed in both scenarios (4a and 4b).

6) Scenarios: action...reaction

The vignette-study also entailed several questions about the possible or most likely reactions in case content is flagged or a user requests its removal. Multiple answers were possible, including giving a warning, blocking the perpetrator for several days, or inviting them to remove the content. The answers given by moderators and OSPs indicate that several reactions are often possible and that reactions are diverse. This corresponds to the subsequent analysis of the technical tools used by OSPs. Furthermore, it is highly interesting that quite some moderators responded with 'I prefer not to answer'. For instance, in case of scenario 1, 6 participants replied with 'I prefer not to answer'. These responses can have several explanations, such as uncertainty, discomfort, or a desire to remain neutral.

In subsample B (OSPs), the responses vary depending on the type of content and the nature of the violation:

- For **OHS** cases, the most chosen response is to 'block for one or more days'. This action aims to provide a temporary suspension as a deterrent and a means to cool down heated situations. However, in our view, this approach may not always address the underlying issues leading to the OHS.
- For cases involving **NCII**, the prevalent approach is more severe. The suggested actions here are to 'remove the image' and 'block indefinitely'. This reflects the serious nature of NCII and the importance of the underlying protected values: such content violates privacy, consent (or sexual autonomy), and is often covered by legal norms. Removing the image helps protect victims' rights and dignity, whereas blocking perpetrators indefinitely prevents them from accessing the platform.

7) Scenarios: comparison with digital natives' assessment in WP3

The results of the vignette-study were compared with the outcomes of WP3 for the same scenarios. This comparison is, of course, limited to the scenarios that were used in both WPs. Moreover, due to methodological constraints, some initially foreseen questions that would be relevant for this comparison had to be deleted from the WP3 survey.

On the one hand, it is worthwhile to note that the permissibility assessment by moderators/OSPs and digital natives does not vary significantly according to the different variables. This is a positive finding because it indicates that moderators'/OSPs' assessment is well aligned with the perception of young people aged 15 to 25.

On the other hand, when comparing the reaction of moderators and digital natives to impermissible online content, both populations seem to be more affirmative and stricter when it comes to NCII than OHS when asked whether the behaviour should be penalised. But there are also differences between both populations. First, with respect to OHS and cultural orientation, the inclination of digital natives towards punishment is more pronounced when the hate target is Black (rather than when it is White), whereas the reactions put forward by moderators remain similar. Second, concerning NCII and gender, digital natives lean more toward content removal when the victim is female (with a 20% increase), while moderators do not make significant distinctions based on gender.

d. Map of technical solutions to remove and prevent illegal/impermissible content

In a last step, the study also aimed to map and analyse the proactive and reactive mechanisms mentioned in the OSPs' policies to prevent the publication of impermissible content or to remove this content if it is published anyway. The mechanisms are categorised in the table below and, more in detail, in the one contained in annex 20, distinguishing between proactive (i.e., before the content appears online) and reactive (i.e., after the content has been posted) tools, between tools put in place by the OSPs and tools used at the request of or by the user of the platform, and finally, between human and automated intervention.

i. General remarks

A number of general tendencies emerged from the analysis of the OSPs' policies. First, policies differ considerably in terms of the amount of detail in the description of the tools used for content moderation, and thus in terms of transparency. Some policies are very explicit about the type of AI tools or software programmes they used to prevent or identify illegal content, and the consequences that follow identification. Others do not mention any tool (BeReal) or only provide a vague description of, for example, 'automated detection and removal' (Reddit), or 'non-public algorithmic and technological tools'. One OSP (Snapchat) even indicates that the way in which the platform is configured limits the possibility of encountering unauthorised content, due to the ephemeral-by-default nature of content. Regarding child sexual exploitation, however, most policies contained references to specific tools on how such content would be identified and removed (e.g., hash-list scanning, image identifiers).

Second, OSPs' approach differs depending on when they were established and the experience they gained as they grew into widely used platforms. Whereas longer established platforms focus more on the tools they put into place themselves (usually a combination of AI and professional moderator teams), it appears that relatively new platforms (turn more towards their users (who can be considered 'non-professional moderators') to implement and sometimes even define the user rules on the platform. On Reddit and Discord, for instance, it will be the administrator or a non-professional moderator of the community (or 'subreddit') or the 'server' who sets the rules (a server on Discord is defined as 'an invite-only home for your friends or community - a place where you can talk, hang out, and have fun'). These rules do not necessarily indicate which content is permissible or impermissible for the platform, but often also reflect personal preferences of the creators (i.e., the persons who create the community) or administrators of the community (i.e., persons who administer the community after its creation; while some administrators are also the creator of the community, others are appointed subsequently) and, in the case of Discord, the server's moderators (who are nevertheless supposed to apply rules laid down by the creators or administrators). All three categories of actors will have the first say regarding content removal. This approach based on users' choices can also have consequences for the use of AI tools. For example, on Discord, depending on the features chosen by the server's administrator, the AI tools will not necessarily remove messages automatically: the bot can also just flag the content and notify the server's moderator; in the meantime, the user can be automatically given a 'time-out'. On Reddit, in contrast, automated tools will still automatically remove content or send warnings to the users.

Third, due to the relatively new development of all tools used by the OSPs, several automated tools to identify, for example, potentially harmful comments before they are posted, are only available in certain languages, sometimes even only in English.

Fourth, automated tools can be used to assist human moderators who intervene proactively or reactively. On the one hand, for example, tools can be used to prioritise review by moderator teams on the ground of the severity, the viral nature and likelihood of content violating certain criteria. AI tools can also be used to select 'borderline' content where it is not clear to the AI whether the content is prohibited (e.g., YouTube) or to detect suspicious items, such as weapons. On the other hand, certain OSPs explicitly state that they use automated tools to reduce the volume of potentially distressing videos that moderators view and to enable them to focus on content that requires a greater understanding of context and nuance.

Fifth, regarding proactive human intervention, not much information can be found in the OSPs' policies. Some policies mention that the automated tools sometimes send the content for a second check to a review team (i.e., human moderators) to decide on the outcome. Interestingly, Instagram's policy explicitly mentions that this closer look by human moderators enables the AI tools to learn and improve their detection capacity (even if it seems obvious that most AI tools today have self-learning capacities). Others refer to a more intensive human intervention after AI tools have flagged the content (e.g., Pornhub indicates that all flagged content is reviewed by trained staff before it goes live on the platform). Only Snapchat seems to provide an active role for human proactive intervention that is not related to the use of AI tools. Its policy mentions that the 'Spotlight' function (which allows users to share content with the entire Snapchat community) is proactively reviewed by human moderators before the content can reach more than 25 people.

Finally, OSPs' policies do not always make a clear distinction between automated tools that are used proactively and those that are used reactively. Presumably, proactive automated tools can often be applied in a reactive manner as well, but we did not always find a clear affirmation of this in the policies. For this reason, we have chosen to include in the column with reactive automated tools entitled in the table below only the AI technologies that the OSPs explicitly identify as reactive tools.

In addition to these general remarks concerning the technical tools used by OSPs, the team also investigated the possible consequences of these content moderation tools. Those consequences impact either the content, the user who posted it, or both. Most of these measures are generally applied by all OPSs, with some specificities (e.g., the number of warnings before an account is suspended). It should also be pointed out that some OSPs provide very detailed information about the possible consequences of moderating actions, depending on the type of behaviour observed and the type of content posted, whereas others simply list the possible consequences without linking them to a particular behaviour or content. Future research would be welcome to provide a more detailed mapping of those consequences of content moderation practises as well as their potentially disproportionate impact on users' fundamental rights.

ii. Mapping of technical tools used to tackle impermissible content

The research continued to map several technical tools according to the relevant criteria, namely (i) whether they are installed by OSPs or user-generated, (ii) whether they are automated or require human intervention, and (iii) whether they are proactive or reactive tools. This mapping includes consequences related to the content and measures related to the user, as developed in the following table (for a more detailed analysis, see the table in annex 20).

TOOLS INSTALLED BY OSPs

CONSEQUENCES RELATED TO CONTENT	● Provide context on sensitive or misleading content, including labelling or tagging content; ● Reduce dissemination of 'problematic' content; ● Restrict access to content (e.g., age restrictions); ● Restrict pages/groups from certain monetization features (which enables users to earn money, for instance, through ads); ● Impose an obligation to have an administrator/moderator approve posts; ● Block or remove content/messages/pages/…; ● Sometimes, suspend content from public view upon receipt of a content removal request (Pornhub) or automatically removed if flagged by trusted flaggers (Pornhub); ● Specific rules for content posted by public figures (e.g., Facebook limits the diffusion of 'problematic' – even if permissible – content posed by a public figure).
MEASURES RELATED TO USERS	● Warnings; ● Restriction on creating content; ● Permanent or temporary suspensions (of access to some features or full suspension of the account); ● Termination of accounts: often after repeated violations, such as the three-strikes rule (e.g., YouTube, LinkedIn); possibly after a notification to give user time to download the data in-app (e.g., TikTok); and ultimately, even when there are objective grounds to reasonably believe that a user is about to seriously breach the terms or community guidelines (e.g., TikTok); ● Specific rules for public figures (e.g., temporary restrictions on Facebook).
USER-GENERATED TOOLS	
CONSEQUENCES RELATED TO CONTENT	● Possibility to lock messages containing specific keywords from being sent and to log flagged messages as alerts for the server's administrator to review; ● Possibility to delete and report (to professional moderator) content.
MEASURES RELATED TO USERS	● Issue of warnings; ● Time-out for users (i.e., they will be unable to send messages, react to messages, join voice channels or video calls but they will still be able to see messages) until the decision of removal; ● Possibility to ban users from the community in point (Discord, Reddit).

Table IV. Technical tools used by OSPs and consequences for content and users

e. General findings

The above results have led us to draw several general conclusions. First, OSPs neither use nor define the term cyberviolence, but prefer to distinguish between various types of impermissible online content and to adopt separate policies depending on the type of content. Each policy defines different permissibility criteria, which vary according to the behaviour observed and the category to which it belongs. This sequenced approach makes the broad phenomenon of cyberviolence more sizeable for OSPs.

Second, when comparing the definition of impermissible content in a sample of twelve selected platforms, the variety in categories, labels and policies is striking. What is permissible on one platform, may be impermissible on another. Moreover, the impact of the legal framework of the user's location on the definition of impermissible content seems to be very limited. OSPs define their own policy rules on impermissible content in cooperation with experts from both academia and civil society, or they try to find the common denominator in the various legal frameworks they are subject to when developing globally applicable policies. Moreover, the borders of permissible online content are evolutive and context-sensitive. All OPSs' representatives involved in the @ntidote project agreed to the fact that their community guidelines are a living document. The OSPs' self-regulatory framework is indeed regularly updated to account for new behaviours. Moreover, the policy rules on impermissible content are written in a somewhat vague or open wording, clearly leaving room for interpretation, which may result in different decisions concerning the same behaviours, whether adopted on the same platform or on different platforms. Consequently, users may not always understand what is (im)permissible content when they use an online platform or be able to foresee the outcome of the content moderation process. The @ntidote project thus confirms the huge power and margin of discretion OSPs have when moderating online content.

Third, confidentiality reigns in the content moderation realm. OSPs are reluctant to provide information on the internal process, largely sticking to the information that is publicly available to users. Whereas such information is quite detailed for some OSPs, others only provide a minimum of details. For instance, it is often unclear whether any feedback is given to the flagger, in addition to the feedback given to the user who posted the content, and there is a lack of transparency on the process that leads to a concrete content moderation decision. It is also very difficult to know, based on the self-regulatory framework, how the technical tools are shaped or controlled to avoid or remedy errors. This lack of transparency is problematic, especially in combination with the aforementioned flexibility left by OSPs' definitions of OHS and NCII. Not only will users not understand why a moderation decision has been taken in relation to the content they posted; they will also not know how the decision was taken. This becomes even more problematic when the moderation process involves non-professional moderators, especially if they receive no specific training and if their decisions are not subject to the OSP's evaluation.
As for individual moderators, they are bound by non-disclosure agreements, which proved to be a huge obstacle for the qualitative research in WP4. Despite extensive recruitment efforts, the sample in the moderators' survey is very limited. That said, the results of the survey on moderators' training, support, and challenges are rather positive, in contrast to much of the existing literature.

Fourth, while one of the research objectives was to get a better understanding of what factors play a role in the permissibility assessment of online content, the results of the survey among moderators and OSPs are limited. Certain variables (e.g., age, consent, cultural diversity) have not or insufficiently been tested. The impact of other factors seems to be surprisingly limited (e.g., gender, sexual orientation, skin colour), even if further research on a bigger sample would be necessary to confirm these indications.

Fifth, different parts of the research confirm that technical (often AI) tools and human moderation go hand in hand. Certain OSPs provide detailed information on the kinds of automated tools they use; others do not mention any tool or only provide a vague description. The use of technical tools and their interaction with human moderators vary depending on the OSP and the type of content.
Technical tools seem to be used primarily to prevent (i.e., proactively) or detect (i.e., reactively) impermissible content; certain OSPs mention specific reactive automated tools. From the moderators' survey, we learnt that the scenarios involving NCII appear to be slightly more prone to detection by automated tools than OHS; this also resulted from the roundtable with industry. Moreover, human

moderators rarely intervene proactively; instead, they react to user notifications or to content flagged by AI tools. The moderators' survey, however, suggests that individual moderators do not always know if the content they review was flagged by an automated tool or a user. Neither do they seem to be sure whether the content would be flagged or not by an automated tool, suggesting the functioning of these tools is a black box to moderators. Furthermore, it is also interesting to highlight that more recently established online platforms more often turn to non-professional content moderators than older, more established, 'first-generation' online platforms; those non-professional moderators may define their own permissibility criteria, in addition to the platform's policy rules.

Sixth, most recently, the EU legislator adopted a new regulatory framework for providers of digital services aimed at enhancing user protection, at providing meaningful accountability of those providers, and at empowering recipients and other affected parties. While it is regrettable that the DSA does not address the lack of an EU-wide definition of illegal (or impermissible) content – for this, other substantive EU legislation would be needed, in addition to existing piece-meal legislation on, e.g., racism, xenophobia, and child sexual abuse – it does impose a whole range of due diligence obligations on OSPs and an obligation to cooperate with LEAs. These obligations vary, depending on the service and size of the provider, but will lead to more transparency on how providers deal with impermissible content. Indeed, it will be interesting to reanalyse OSPs' self-regulatory framework in a few years to see whether the current problems regarding the content moderation process will have been remedied or whether further regulation is necessary. For instance, service providers' responsibility to ensure respect for users' fundamental rights is worded in broad terms and the DSA does not entail a general obligation to put in place an external complaint system.

3.5 COPING MECHANISMS & VICTIM SUPPORT

3.5.1. METHODOLOGY

Looking at the qualitative and quantitative data on OHS and NCII, digital natives are clearly regularly confronted with cyberviolence, more specifically for this research with OHS and NCII. After the study of prevalence and further understanding of these forms of cyberviolence among adolescents and emerging adults, the team focused on what actions adolescents and emerging adults take when victimised by NCII and OHS and who they turn to for further support.

First, the research focused on the perception of digital natives based on a survey among the relevant population. Questions about experienced emotions and certain coping mechanisms of victims of OHS and NCII were included in the research's survey on the prevalence of OHS and NCII (see 3.3) and analysed in this section. The same recruitment strategy and procedure is implemented here. Respondents who indicated to have been a victim of either OHS (independently of the type of hate speech), or the non-consensual dissemination of intimate images, were presented with items of the Cybervictimisation Emotional Impact Scale (CVEIS) to measure the emotional impact on the victims (Durán & Rodríguez-Domínguez, 2023). If a respondent was both victim of OHS and of the non-consensual dissemination of intimate images, this scale was presented twice. Durán and Rodrígez-Domínguez (2023) used this scale to measure the emotional impact on women receiving an unsolicited dick pic. As to practical restrictions, i.e., the length and the duration of the survey, the team did not include supplementary questions to measure the psychological or physical impact of victimisation more in-depth.

As stated above, to try to reduce the harm caused by victimisation, a victim's coping strategy can be to ask advice or help from several sources in their environment. The team decided to include questions that map what sources, and to which extent, play a role in coping with victimisation. The answering options were based on the input of the interviews of WP1 and on literature concerning coping with related forms of victimisation (e.g., offline sexual abuse) (Bal et al., 2009; Garcia, 2010; Guerra et al., 2018; Margaret et al., 2018; Valido et al., 2020). As to practical reasons, i.e., the length and the duration of the survey, the team did not include validated scales that cover other elements of coping, such as denial and substance abuse. If a respondent was a victim of both OHS and the non-consensual dissemination of intimate images, the same question on coping was presented twice. Table IV (annex 10) presents the questions measuring the emotional impact and advice seeking behaviour of victims.

Second, the research focused on the experiences of those organisations that are particularly tasked with addressing OHS and NCII by applying qualitative research. For these interviews Unia, the Institute of the Equality of Women and Men (the Institute), and Child Focus were selected based on their current role as focal organisations, established by law, for receiving complaints on OHS (Unia) and NCII (the Institute for adults, Child Focus for minors) and acting *de facto* as trusted flaggers vis-à-vis O SPs for the removal of content. In addition, given the particular focus on LGTBQIA+ as a particular subgroup in the research, the organisation Çavaria was also interviewed for additional input on OHS and NCII. Other support organisations were contacted but declined to be interviewed due to time constraints and lack of capacity among their team. The interviews were conducted as semi-structured interviews, lasting from 60 to 90 minutes. Within the time constraints, the Institute for the Equality of Women and Men preferred a written reply to the questions. This part of the research is explorative in nature in that it intends to include the expertise and insights of those organisations that work with victims of OHS and NCII on a daily basis, in order to test the team's findings and recommendations.

3.5.2. RESULTS

a. Adolescents' and emerging adults' perspective: emotions and coping

In WP5, the team focused in the first place on the emotions and coping mechanisms of the victims of OHS and NCII. The literature study shows that very limited research has been done in this field as of yet. In the survey, respondents who had replied positively to questions of victimisation regarding OHS and NCII were questioned about their emotions upon the confrontation with these behaviours of cyberviolence.

i. Negative emotions experienced by victims of OHS and NCII

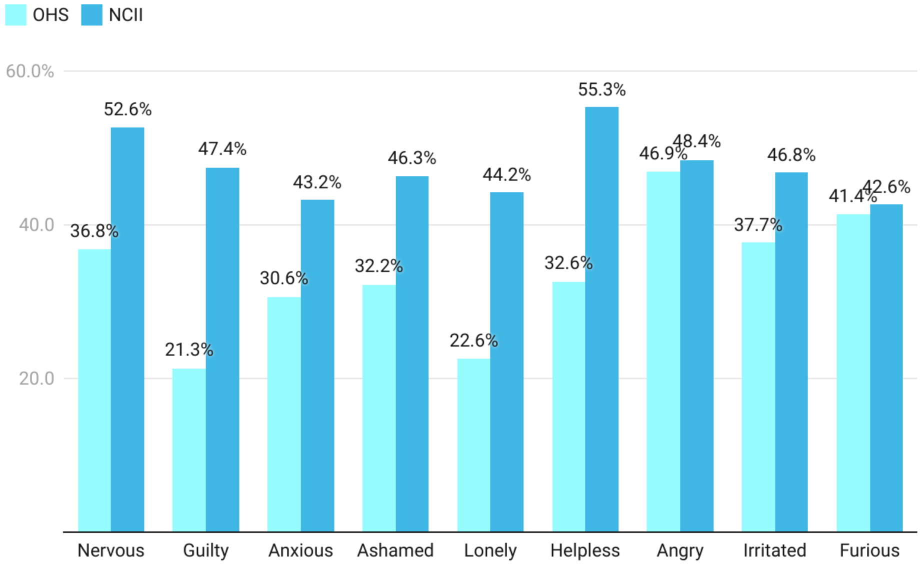


Figure 25. Negative emotions in victims

The results presented above show that victims of NCII experience in general more negative emotions than victims of OHS. Also, the feelings most encountered by victims are different for the two behaviours of cyberviolence.

Among the respondents, the most common feelings after victimisation of OHS are feeling angry (46.9%) or furious (41.4%). An explanation for the high level of anger among victims of OHS could be the fact that this behaviour targets the personal characteristics which form their identity, such as their nationality, sexual orientation, ethnicity, or gender. Victims of OHS are often (as stated in WP3) targeted because they belong to a minority group in terms of gender, ethnicity, and sexual orientation. Further, one third of the victims of OHS feels nervous (36.8%), ashamed (32.2%), helpless (32.6%), and irritated (37.7%), while one out five victims feels guilty (21.3%) and lonely (22.6%). Feelings of helplessness and loneliness might be explained by the fact that victims feel that they are among the few that encounter such an event. Moreover, the often-reported high levels of OHS could enhance the feeling that nothing can be done about OHS, which might lead to feelings of loneliness and helplessness.

For NCII, the most reported feelings associated with victimisation are nervousness (52.6%), helplessness (55.3%), anger (48.4%), guilt (47.4%), and irritation (48.8%). Further, more than 40% of NCII victims feel anxious (43.2%), ashamed (46.3%), lonely (44.2%) or furious (42.6%). Feelings of helplessness, anxiety and nervousness can be explained by the fact that victims cannot control to whom their picture is being disseminated or the impact of this dissemination, and cannot directly take these pictures offline themselves. Victims of NCII often feel ashamed and guilty because they think it is their fault the intimate image has been disseminated, as they "should not have sent it in the first place", which is so-called victim-blaming. Feelings of anger and irritation can be explained by the idea that someone broke their trust by sharing such private information.

ii. Victim support sources

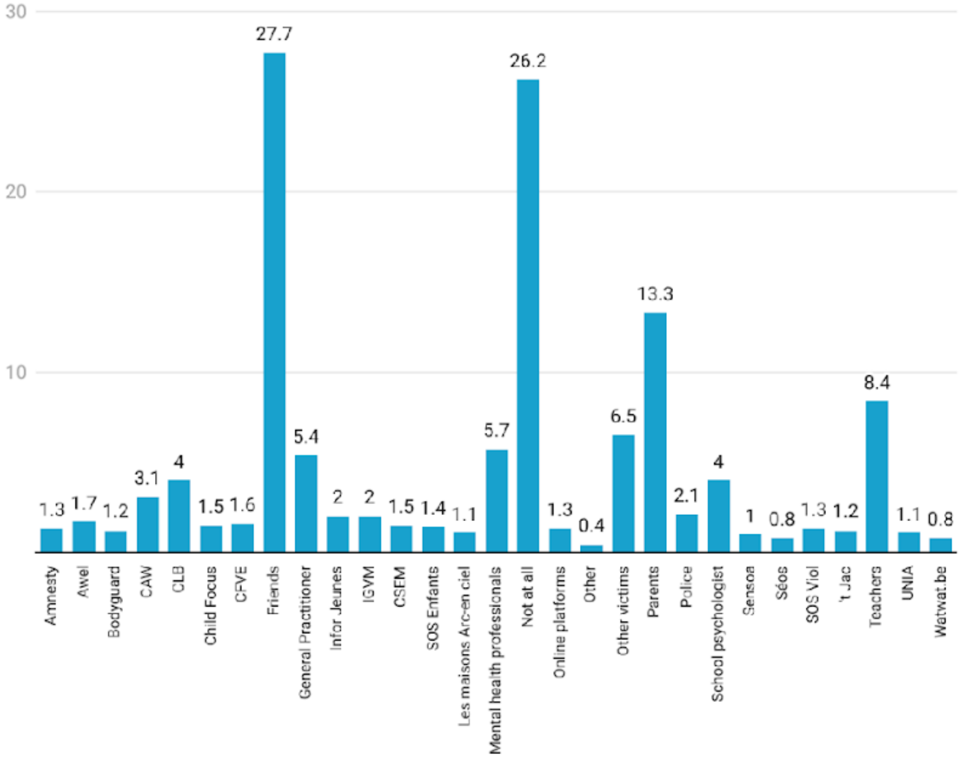


Figure 26. Victim Support Sources OHS

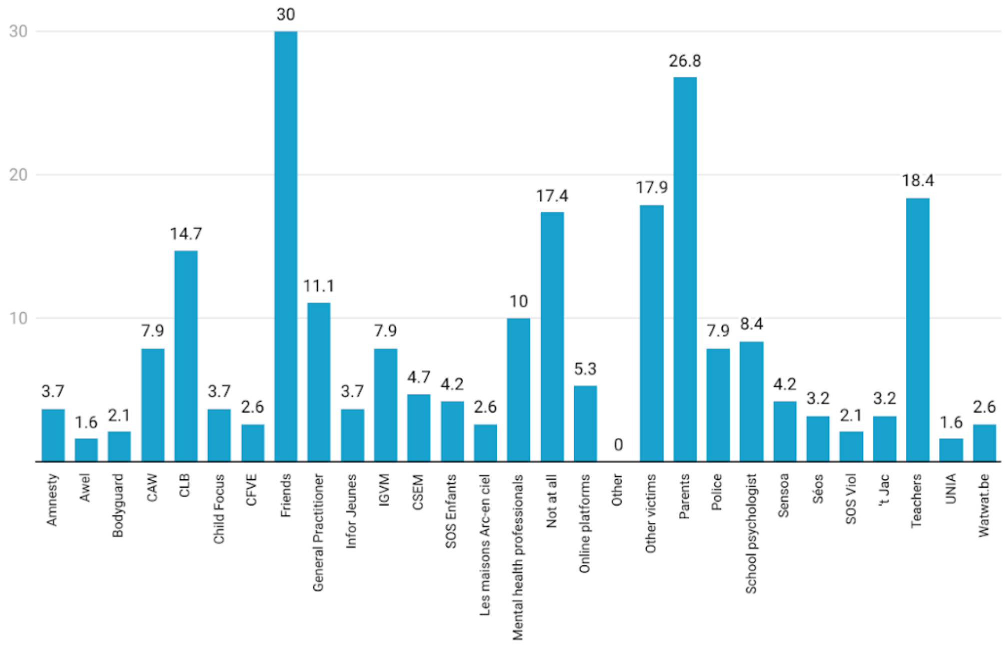


Figure 27. Victim Support Sources NCII

One of the coping strategies to deal with victimisation is to reach out to external resources for support. Overall, NCII victims reach out more to support sources than victims of OHS. This can be explained by the sensitivity and the intimate character of the event. Moreover, almost half of NCII victims experience negative emotions after being victimised and as such take the logical step to reach out to sources of support. In both types of victimisation, the results show that less than 10%, and in the majority of cases even less than 5%, of the victims reach out to official support sources (i.e., Amnesty, Awel, Bodyguard, Infor, Institute for the Equality of Men and Women, Le Conseil Supérieur de l'Education aux Médias, Les équipes SOS Enfants, les maisons Arc-en ciel, Sensoa, Service d'Ecoute d'Orientation Spécialisé (Séos), SOS Viol, 't Jac, UNIA, watwat.be). Although almost all these victim support sources are equipped to support victimisation of OHS and NCII, digital natives scarcely rely on these sources. A first reason may be that adolescents and emerging adolescents do not know these sources exist and consequently do not contact them after victimisation. Secondly, victims might not reach out because they experience emotions of guilt and shame. A third explanation may be that, given the high levels of helplessness reported by victims of NCII and OHS, they might believe nothing can be done against these behaviours and their impact.

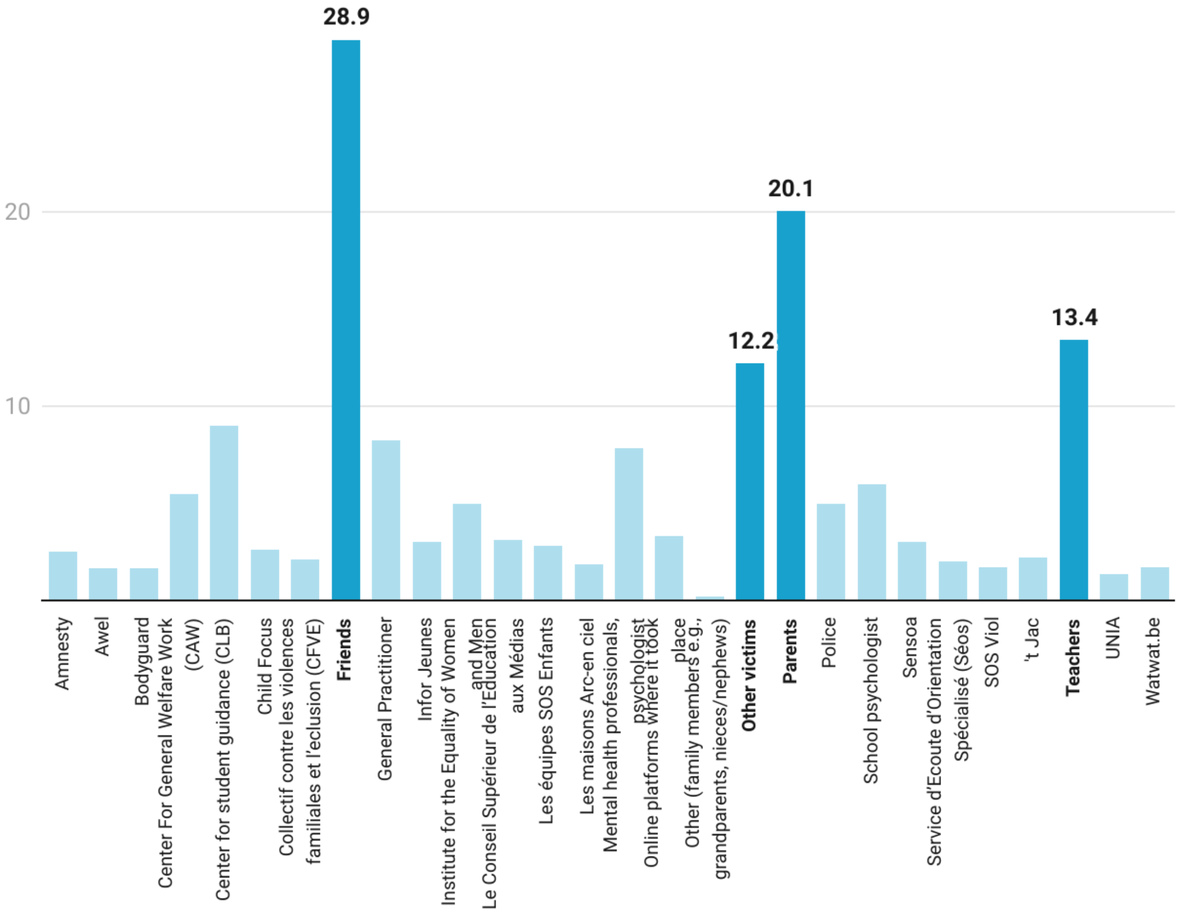


Figure 28. Mean of OHS and NCII

The majority of the victims reach out to their direct environment, including both informal and formal networks. Both NCII and OHS victims primarily reach out to friends ($M_{friends}$= 28.80%) as they are more present in their daily life activities (informal network). Victims also reach out to teachers (Mteachers= 13.40%), their parents ($M_{parents}$= 20.05%), and other victims ($M_{othervictims}$ = 12.20%). The prevalence rates of NCII victims asking for help from their parents and teachers doubles in comparison to victims of OHS. This might be explained by the sensitivity of the event. On average, 10% of the victims also reach out to their formal network (general practitioner, psychologists, school psychologist). The results show that the prevalence rates of NCII victims asking help from their formal network doubles in comparison to victims of OHS.

On average, only 5% of victims contact the police, while the majority of victims considers both behaviours as harmful (see WP3). This might be caused by the feelings of shame they experience or because they are not aware of the fact that these behaviours are in fact criminal and punishable by law.

b. Organisations' perspective

In the following paragraphs, we present some main findings on the role, perceptions and challenges of the interviewed support organisations.

General findings

Upon analysing the interviews, the team found certain similarities in the replies of the organisations relevant to both OHS and NCII, namely regarding (i) the organisational framework of support for OHS and NCII (i.e., role and capacity), (ii) cooperation with other stakeholders, (iii) the specific role of these organisations as *de facto* trusted flaggers and the cooperation with OSPs, (iv) the interaction with minors and early adults, (v) the challenges posed by digital technologies, (vi) the support needs of victims, and (vii) focus on cultural change. Specific for OHS, the legislative framework was of particular concern (viii).

i. The organisational framework of support for OHS and NCII

The support organisations generally indicate that they focus on three areas, namely providing direct victim support (online and/or via telephone, i.e., so-called hotlines), focusing on preventive work by advising stakeholders such as schools, parents, sport clubs or other partners, and focusing on policy work and advising authorities. The organisations, therefore, are focused on specific issues of cyberviolence, e.g., OHS or NCII, and/or specific focus groups of potential victims, e.g., gender, age, or sexual orientation. For example, whereas Child Focus will support victims of NCII who are minors, IEWM will focus on victims who are adults. Likewise, where UNIA will address OHS based on all grounds (skin colour, race, or gender), IEWM will in addition to NCII support victims of OHS but only from a gender-based perspective.

The organisations with hotlines for victims all indicate that they receive a significant number of complaints and questions but do not have the staff and resources to adequately respond to all these requests (annex 21 with numbers for Unia and IEWM). As such, several organisations indicate that they need to prioritise certain requests over others, e.g., one of the organisations indicates that they internally decided to focus on complaints of victims and not proceed with complaints made by bystanders. Another organisation argued that a sizeable amount of their budget was project-based, meaning that at the end of such a project, the means also disappear to provide prolonged and steady support of victims.

ii. Cooperation with other stakeholders

All organisations highlight the high level of cooperation between the organisations included in the study, other organisations focusing on these behaviours, and other stakeholders, such as schools or police. In case of overlap, the organisations state to cooperate with each other. Also, where an organisation might not have all the instruments for supporting victims in comparison to another support organisation, cooperation appears to complement such gaps. For example, Çavaria will refer victims of OHS who are considering filing a complaint to Unia for further support, in line with an established practice between both organisations. Also, Sensoa cannot file a criminal complaint for sexual violence, but might refer victims to IEWM.

While the organisations are generally positive about cooperation between them, they also highlight the potential for duplication of programs and lack of coordination on a more structural level. One organisation made the connection with the focus of policy on project-based budgets for the organisations, which stimulates short term projects that are not structurally imbedded within a stronger framework of cooperation.

All organisations highlight the importance of cooperation with schools. Several have developed specific material for schools or actively cooperate with specific schools on themes relevant to OHS or NCII (e.g., 'transgressive sexting' or consent online). During the interviews, all organisations put focus on the importance of schools and formal education as essential in preventing OHS and NCII. After the finalisation of the interviews, an open letter by several representatives of certain organisations, including several that took part in the interviews, highlighted how the lack of focus on relational and sexual health in the new compulsory teaching terms in secondary education undermined efforts to prevent transgressive sexual and relational interactions (Magits, 2023). Moreover, previous research by the Flemish centre of expertise in media literacy showed that cyberviolence, including online image-based sexual abuse, is scarcely addressed in primary school and not addressed in a significant share of the schools (Mediawijs, 2022).

iii. The roles of the organisations as trusted flaggers to OSPs

Of the organisations, three organisations (Unia, IEWM and Child Focus) act as *de facto* trusted flaggers vis-à-vis OSPs, meaning that victims or bystanders can report OHS to Unia and IEWM (sexist OHS) and NCII to Child Focus and IEWM. Child Focus and IEWM are also accepted as trusted flaggers by some OSPs, meaning that reports from these organisations will be given priority. Moreover, both Child Focus and IEWM are considered partners in platforms set-up in cooperation with the industry and hotlines that function as trusted flaggers for the removal of NCII, namely stopncii.org focusing on NCII of adult victims and #Takeitdown for CSAM, which may include NCII of underage victims. Other organisations that do not function as trusted flagger regarding OSPs, such as Çavaria, indicate to closely cooperate with Unia and IEWM in such cases.

All three organisations highlight that they regularly contact OSPs for removal of content, albeit Child Focus and NCII on a regular basis for NCII whereas Unia does so less regularly for OHS. Unia explains that they will only report OHS to the OSPs when they consider the speech unlawful in view of the Belgian legislation, even if the threshold in the terms of services of the OSPs is lower. Unia reports that, even given this particular care and self-restriction when reporting content, posts are infrequently and incoherently removed depending on the given OSP and the (changing) policy. In some cases, Unia does not even receive a reply to a takedown-report. IEWM and Child Focus tend to be more positive on the cooperation with OSPs for the removal of NCII and notice a positive trend.

However, both indicate that still too often NCII is removed only after some time and some OSPs are more responsive than others. IEWM highlighted the importance of this cooperation, as removal of images is often the first question asked by victims who reach out to them.

Until now, such cooperation was *ad hoc* and based on the willingness of the OSPs to cooperate. However, during the research the DSA entered into force, including the recognition of 'trusted flaggers'. Article 22 of the DSA provides for a status of trusted flaggers for those national organisations that can receive complaints of illegal content on online platforms and provide assistance to those who submit a complaint or are victim of the potential online content and conduct, and who notify platforms of the illegal content on their sites. National organisations can be accorded this DSA 'trusted flagger' status after an application at the digital services coordinator of the Member State. The consequence of the trusted flagger status is that online platforms need to take the necessary technical and organisational measures to ensure that reports submitted by trusted flaggers on illegal content, in this case OHS and NCII, are given priority.

The main criteria for being accorded the status of trusted flagger under the DSA are that the organisations must have particular expertise and competence for the purposes of detecting, identifying, and notifying illegal content, that they are independent from the OSPs, and that they carry out their activities for the purposes of submitting notices diligently, accurately, and objectively. From the interviews of Unia, Child Focus and IEWM, it appears that all will qualify as candidates for the status of trusted flaggers, continuing their current informal role in relation to the OSPs. With the DSA, they will have more leverage towards OSPs to remove illegal content.

iv. The interaction with minors and emerging adults

The interviews also show that several organisations are not often contacted by minors (< 18 years) themselves. This might not come as a surprise for IEWM, for example, which focuses on NCII where adult victims are concerned, but also for Child Focus, which indicates that they are mostly contacted by parents or other adults regarding concrete cases of NCII (or broader, CSAM). The same goes for OHS, as Unia also reports not to receive many reports from minors or emerging adults. The organisation suggests that this might be due to adolescents and emerging adults not knowing them as well as adults and/or a higher level of acceptance of 'harsh language' among them.

This finding translates back to the results from the survey that show that digital natives in such cases will first contact friends and other adults and will only in a limited number of cases contact support organisations directly. Reaching out to the adult support network, e.g., parents or teachers, might result in contacts with the support organisations, either for filing complaints or for asking advice on how to deal with the victimisation of the minor by OHS or NCII. However, there also remains a group of victims that simply did nothing.

v. Challenges posed by online technologies

Organisations highlight challenges posed by online technologies on fighting OHS and NCII. The most important challenge is the wide variety of OSPs with their own terms of services and ways of reporting. Child Focus highlights the importance of cooperative platforms, where several OSPs cooperate with trusted flaggers to remove an intimate picture. However, organisations are aware of platforms that are simply stonewalling and as such cooperation is impossible.

The organisations indicate that the question of removal has become more difficult due to the fact that OHS and NCII are ever more often disseminated via so-called ‘private socials’, including instant messaging systems, Discord boards, Snapchat or DMs on social media. As such, the majority of NCII and OHS is disseminated one-to-one or in private groups, limiting the potential of finding and removing such content. The organisations further warn of the impact of AI on OHS and OSPs, including rampant OHS in the Metaverse or the use of AI for deepnuding.

The organisations also highlight that technology could be used to tackle cyberviolence. However, Child Focus mentions that there appears to be more willingness by OSPs and other tech companies to develop and apply instruments for the removal of CSAM than NCII, e.g., tools for the hashing of intimate images, using photo DNA for retrieval and removal of such images, or the exchange of image databases are more frequently used in cases of CSAM than NCII. The application of such technology to only certain forms of unlawful content, as well as the previous experience of the organisations that CSAM is more easily removed than NCII, which is in turn more easily removed than OHS, suggests an informal hierarchy among the OSPs in scanning and removal. At the same time, several organisations also highlight the promise certain technology holds for automatic finding and removing of content, even though this technology is not yet sufficiently developed and applied.

vi. The support needs of victims

The organisations highlight that they are most often contacted for questions of advice, information, and removal of content (trusted flagger-function). Particular to NCII, Child Focus and IEWM highlight that the question of removal of content or prevention of dissemination is the first objective of victims when asking for additional support. As is also clear from the case law study in part 3.2, victims of NCII are mostly motivated by removal of intimate images or preventing dissemination rather than criminal action to find and punish a perpetrator. Another reason for contacting these organisations is to get additional information on how to discuss OHS or NCII with adolescents and emerging adults (e.g., by parents, schools, or others).

In particular with regard to NCII, the relevant organisations highlight the importance (and often lack) of psychological support. Given the high impact of NCII on victims, as is clear from the survey on emotions, this does not come as a surprise. However, most organisations are able to refer victims to primary care where victims receive a low amount of therapeutic sessions. Afterwards, they are often on their own for a while due to the long waiting lists to receive consistent and long-term therapeutic intervention.

vii. The focus on cultural change

Several of the organisations highlighted the importance of ‘cultural change’ as to cyberviolence in order to tackle OHS and NCII. Several organisations also remarked ‘harsher’ language being used and seemingly tolerated online, e.g. Unia argued that they perceive a higher tolerance for hate speech online, which they also believe to account for a lower level of complaints in comparison to other forms of discrimination and hate. Another organisation argued that increased polarisation in society results in an increased (online) presence of anti-LGTBQIA+ groups and sentiments.

As such, cultural change as to what is acceptable and unacceptable online is crucial for tackling OHS. Child Focus highlights that the same is true for NCII, where there is a need for a better debate on (online) consent and boundaries. Several of the organisations in this respect developed material for schools or other stakeholders to develop a better notion of consent and boundaries in the virtual arena. Child Focus argues that, while the general messages on boundaries should be frequently repeated, there is also a need for more targeted communication and material for certain groups, e.g., material for boys on toxic masculinity.

viii. Legislative framework of OHS

Whereas most organisations argue that a developed legal framework for NCII is generally provided, several organisations mentioned the procedural hurdle of article 150 Constitution for prosecuting unlawful hate speech as a major obstacle in tackling OHS. In addition to hampering effective prosecution, it was also argued that the differentiation between on the one hand racial and xenophobic OHS and on the other hand OHS based on other grounds is out of odds with current forms of OHS. Several organisations remark that OHS is often focused on several characteristics. Unia remarked an increase in OHS based on religion in combination with other grounds of racism and xenophobia. Given that current OSPs are rather unlikely to cooperate for the removal of OHS, it appears that for certain forms of OHS (e.g., homophobic or sexist), there is no effective remedy provided for victims.

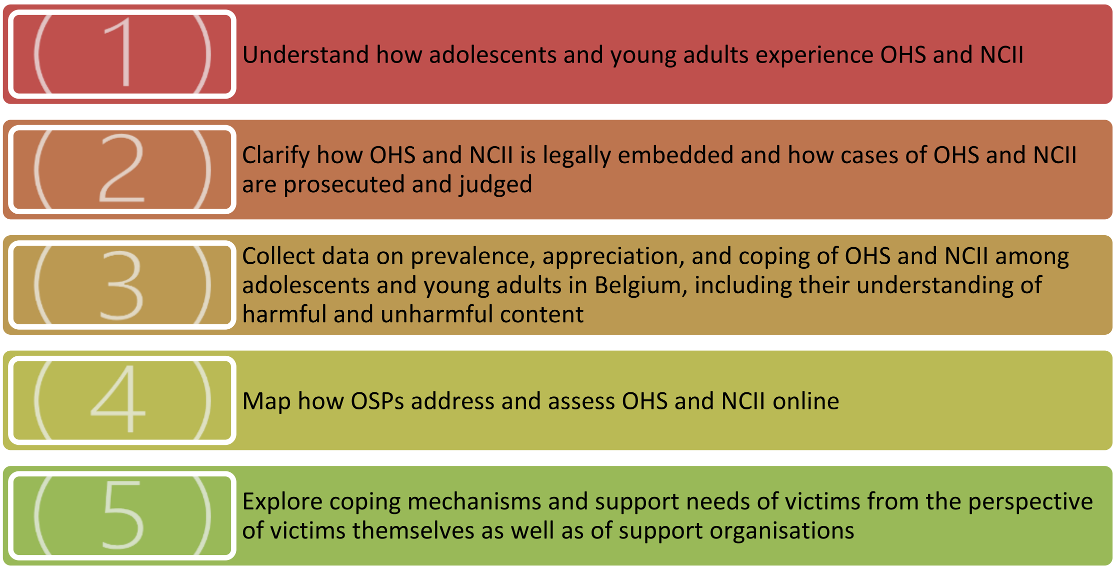
c. Conclusions

The study shows that while there are support organisations with knowledge on and expertise in OHS and NCII present in Belgium, there remain budgetary, capacity and legislative constraints that prevent these support organisations from fully playing their role. A short-term goal in this respect is to ensure that the current hotlines for OHS and NCII (Unia, Child Focus and IEWM) are recognised as trusted flaggers under the DSA. This will further give them leverage to ensure the (quick) removal of unlawful OHS and NCII and play a role in the further development of cooperative platforms of industry and trusted flaggers. Moreover, the absence of structural mainstreaming of knowledge on cyberviolence in the educational curricula and the absence of a long-term vision on a structural cooperation between education and support organisations hinders prevention. The main challenges to improve the effectiveness of victim support for adolescents and young adults is to enhance the visibility of support organisations, specifically for OHS and NCII, increase their capacity, particularly regarding psychological help, and lower the threshold to seek support.

4. GENERAL CONCLUSIONS AND RECOMMENDATIONS

4.1. GENERAL CONCLUSIONS

The @ntidote project started from the assumption that adolescents and young adults in Belgium are regularly confronted with cyberviolence, particularly online hate speech (OHS) and non-consensual dissemination of intimate images (NCII). Previous research, both in Belgium and abroad, has signalled that the omnipresence of social media and communication apps in the lives of adolescents and emerging adults also resulted in them being regularly exposed to these harmful online behaviours. Therefore, the @ntidote project set out five objectives to better understand OHS and NCII within the Belgian context:



(1) Understand how adolescents and young adults experience OHS and NCII

(2) Clarify how OHS and NCII is legally embedded and how cases of OHS and NCII are prosecuted and judged

(3) Collect data on prevalence, appreciation, and coping of OHS and NCII among adolescents and young adults in Belgium, including their understanding of harmful and unharmful content

(4) Map how OSPs address and assess OHS and NCII online

(5) Explore coping mechanisms and support needs of victims from the perspective of victims themselves as well as of support organisations

Figure 29. general objectives of the @ntidote study

Qualitative understanding of online hate speech and NCII

The study shows the importance of establishing a common understanding among adolescents and emerging adults to avoid misconceptions. Whereas there appears to be a general understanding of OHS and NCII, the study shows that it is crucial to focus on the nuances of vocabulary, as they can have multiple implications, both for recognising the status of victims and for discussing the responsibility of the perpetrators of these behaviours. Also, the same is true for bystanders, who are widely present in the sample and are often combined with another status (e.g. perpetrator or victim). In addition, the results showed that the majority of the sample experienced both online hate speech and NCII. Therefore, it is necessary to consider and adapt the vocabulary that is understood and used by the sample, and to obtain valid responses that capture the nuances of these behaviours. As such, the team learned to take into account the importance of vocabulary use and to concentrate more on this type of methodological precautions. In particular when formulating interview guides and questionnaires. The emphasis lies in fostering genuine engagement with the participants, rather than projecting the researcher's own knowledge, beliefs, or viewpoints onto them. The team strongly advocates to involve digital natives more in ongoing research, encompassing a wide array of profiles, as exemplified by the current research approach.

Moreover, in terms of prevention, it is necessary to inform people about the various types of motivations (based on the motivations perpetrators perceive) to disseminate online hate content and NCII. Social motivation is not the only one discussed. Digital natives highlight the presence of immaturity and emotional effect as underlying these cyberviolence behaviours. In addition, the intentional motives were also present in both behaviours. In conclusion, various perceived motives may exist regarding NCII and OHS and must be considered in preventive actions by providing information about the reasons behind perpetration or to implement in a more forensic care setting with perpetrators. Even if there is a bystander-perpetrator overlap, the motivations linked to bystanders seem to be an area of research to further develop.

The study further demonstrates that victims tend to turn more towards their circle of friends rather than their family or formal institutions. Like in other studies (Lee et al., 2019), the team can speculate this can be explained by the fact that young people have a higher digital knowledge compared to other age groups. Thus, it is necessary to (i) continue media education for parents, teachers, and all other relevant individuals, and (ii) integrate peers into certain awareness-raising actions. Furthermore, victims have also emphasised isolation, which is sometimes chosen to avoid creating a negative impact within their family. The digital realm is rarely used as a resource when respondents directly discuss coping mechanisms. However, respondents also mention the benefits of virtual networks. Therefore, it would be useful to (i) focus on preventing the feeling of isolation, (ii) inform about the underreporting of complaints, thereby initiating a discussion about the lack of prosecution of certain behaviours to increase the legitimacy of initiating legal procedures, and (iii) identify the digital domain as a resource in itself (and not only a threat), notably by communicating on the ease, cost-effectiveness, and anonymity of platforms for the benefit of the victims.

Related to NCII, it would be beneficial to focus on prevention efforts for consensual intimate image sharing, emphasising the roles, responsibilities, and boundaries of each partner involved. It is necessary to question the responsibility or lack thereof of the individual who shares their own photo and to identify the potential implications of this type of sharing. Our results are nuanced, but it would be interesting to initiate a discussion about the role of responsibility and the consequences for the perpetrators, victims, and bystanders. Therefore, it appears necessary to (i) raise awareness about the reasons that can drive a person to send intimate images of themselves to someone else, and (ii) initiate discussions with young individuals about the individual responsibility of each person, including those who are bystanders of NCII. From the perspective of potential perpetrators, it becomes important to identify the harmful aspects that arise from the non-consensual sharing of such photos. Additionally, discussions are needed about the role of bystanders, who find themselves caught between their own emotions and the prevailing social norms within the group where the images are shared (Harder, 2021). The team has indeed highlighted within the literature the role of rape myths in shaping perceptions of victim accountability and supporting the perpetrator (Dekker et al., 2019). Thus, more broadly, it is necessary to better understand the mechanisms of communication and regulation among peers online within discussion groups regarding NCII as some of our respondents shared that NCII were spread through this way.

Related to OHS, the results suggest reevaluating the definition provided encompass both the intended objectives and the actual consequences of such speech. This broader definition would consider the personal interests and individual targeting observed in the context of hate speech. Therefore, it seems essential to explain and inform about the presence of aggressive and hateful messages online, independent of their classification as hate speech (aggressive or hateful).

Indeed, researchers advise against adhering to an overly restrictive vision of online hate speech (Perry, 2001; Schweppe & Perry, 2022), which is also reflected in the definitions provided by respondents of WP1. Thus, the team invites discussions on the treatment and preventive and legal aspects of various nuances within behaviours, considering the diverse motivations and definitions highlighted. Secondly, it would be interesting, both in terms of criminal policies and prevention, to consider the various motivations. Indeed, reducing the behaviour to the produced content tends to overlook the intricacies of the behaviour, and therefore ends up being far removed from the reality experienced by individuals aged 15 to 25. Finally, specificities within the target group are observed, particularly linked to identity development. The digital context appears to magnify the relationship with identity (Keipi et al., 2017), whether it's about self or others (Tajfel, 1979). Therefore, it seems essential to establish appropriate media education programs, including offering platforms for discussion where young individuals can engage in conversations about identity aspects both offline and online.

Regulatory framework mapping of OHS and NCII

The legal analysis makes it clear that there is a wide range of norms enabling the prosecution of OHS and NCII at the national level. Whereas the national legal framework on OHS is supported by a well-developed set of rules at the international level, the level of the COE and the EU level, the international and supranational legal framework of NCII is still under development. The research reveals that the legal framework of OHS is underpinned by the notions of equality, freedom, democracy, and human dignity. The research also shows that, among the legal framework, there is a particular priority for racism, xenophobia, and gender-based hate speech in the international and supranational legal framework, with less focus on norms concerning hate speech on other grounds. The legal framework on NCII is in turn supported by the principles of equality, the prohibition of gender-based violence and sexual integrity. The research further shows that there is an overlap between norms explicitly targeting (O)HS and NCII on the one hand, and more generic norms that were not particularly drafted to tackle these behaviours but are applicable to manifestations of OHS and NCII, on the other. This overlap can be of added value to tackle OHS and NCII but can also be problematic. For instance, the lack of guidance in addressing the dissemination of images of minors among each other without consent as NCII or CSAM may result in unwanted consequences.

Whereas the legal analysis demonstrates that from a normative side law enforcement and courts are well-equipped to address complaints and cases of OHS and NCII, the coding exercise shows that (i) there were only a limited number of complaints compared to the prevalence of OHS and NCII suggested by the literature and WP 3, and (ii) that the vast majority of cases is discontinued on a wide variety of grounds. Only a handful of cases will end up in the courts. For OHS cases, the research discerned ‘clusters of hate’, i.e. the finding that OHS often contains several grounds of hate speech (e.g., comments targeting both the religion, skin colour, and the gender of the victim), suggesting that the prioritisation of certain forms of hate speech is ill-fitted with the actual forms of hate speech. For the NCII cases, the team found that there is a particularly high prevalence of cases either constituting either sextortion or constituting intimate partner tech abuse. This suggests that victims will be particularly motivated to file a complaint in the presence of elements such as financial loss or intimate partner violence, in addition to NCII. Victims of NCII are particularly focused on seeing the dissemination of their images stopped, removed, or prevented rather than on prosecution and punishment of the perpetrator. Furthermore, for both OHS and NCII, the high level of discontinuation of complaints are due to, e.g., a lack of capacity and prioritisation by LEAs. Specifically for OHS, the procedural hurdle included in article 150 of the Constitution whereby OHS, with the exception of racist and xenophobic OHS, needs to be prosecuted before the Court of Assize, has a serious impact on prosecuting these cases.

Prevalence and perspective on OHS and NCII among digital natives

The prevalence study into OHS and NCII shows that there is only a significant difference found between gender subgroups in NCII perpetration: men, transgender people and non-binaries appear to disseminate an intimate image more frequently. Although extensive previous studies established that women become a victim more often than men, this result is not reflected in the @ntidotestudy. The team did not find any significant differences between men and women. Other criteria are significantly related to OHS and/or NCII, namely sexual orientation, ethnicity, and age.

First, in terms of sexual orientation, significant differences are present for the victimisation of specific types of OHS, namely based on gender and on sexual orientation, i.e. more members of the LGBQTQIA+ community receive gender based OHS and OHS based on sexual orientation in comparison to heterosexuals. **Second**, both for OHS and NCII victimisation and perpetration, ethnicity plays a significant role. Individuals with a foreign background, independent of if they were non-Belgian or Belgian, reported to have been more victim of NCII and OHS. Also, for the perpetration of both behaviours, a significant difference between ethnicity groups was detected. An explanation could be that cultural beliefs and values influence how harmful individuals think OHS and NCII are. **Third**, the study reveals that age is a determining factor in the victimisation and perpetration of both behaviours. Emerging adults tend to have been more victim and perpetrator than adolescents. As for victimisation, it could be hypothesised that emerging adults become more often victims of OHS and NCII as they are still developing their identity. This includes the capacity to set their own boundaries and as such, perpetrators can take advantage of the fact that young people are not able to this yet Moreover, significantly more emerging adults are perpetrators in comparison to adolescents. At this point in time, there is more information present on social media regarding coming out as a member of a minority group), tolerating and accepting minority groups, the importance of (explicit) consent. As adolescents are the ones who have been the most exposed to and raised with the idea of digitalisation, their conceptualisation of what is normal and what should be tolerated in a society can differ from what emerging adults have learned.

Overall, the vignettes describing cases of OHS and NCII were both seen as harmful behaviour (>65%). Although there were no significant differences found between the subgroups based on gender, sexual orientation, ethnicity, and age in the vignettes of OHS, ethnicity played a significant role in one set of NCII vignettes (i.e., victim is heterosexual versus LGBTQIA+). People with a foreign background, both Belgian and non-Belgian, think NCII is more harmful when the victim is heterosexual (versus LGBTQIA+). This might underpin that there exists a different conceptualisation of what is harmful and can be explained by for instance differing cultural norms. Age and sexual orientation do play a significant role in the assessment of what is considered as an appropriate legal reaction to OHS and NCII. However, the main conclusion to be drawn here is that adolescents and emerging adults opt more often for an alternative way of sanctioning (e.g., mediation, damage compensation, community service and following an online course) rather than for traditional criminal punishment, such as imprisonment.

Finally, a PWM was built for both behaviours by applying structural equation modelling. For OHS, it showed that someone's intention to engage in OHS is mostly driven by how others (dis)approve of their behaviour (e.g., subjective norm). Someone's willingness, on the other hand, is driven by having a positive attitude towards the OHS perpetrator and perceiving themselves as similar to the perpetrator (i.e., prototype favourability and similarity).

In conclusion, engaging in hate speech is something that occurs when applying the reasoned path as the behaviour was related to intention only. This means that individuals weigh more the advantages against the disadvantages and analytically reflect before engaging in OHS. For NCII, willingness was related to both prototype similarity and favourability: intention was related to subjective norms. NCII is as such a behaviour that is influenced by the social reactive path. How much you think you look like a perpetrator of NCII or having a positive attitude towards NCII perpetrator plays a role in engaging in NCII.

Self-regulatory framework and understanding of OHS and NCII

The team's research on OSPs' self-regulatory framework and understanding of cyberviolence shows that OSPs neither use nor define the term cyberviolence. They prefer to distinguish between various categories of impermissible online content and adopt separate policies depending on the type of content for which distinct permissibility criteria are defined. Those policies are a living document: they evolve in line with new behaviours observed on the platforms. But usually they do not take into consideration the legal framework of the user's location on the definition of impermissible content. Moreover, those policy rules are written in an open wording, leaving considerable room for interpretation, in contrast to the detailed internal rules to be applied by moderators. OSPs therefore enjoy a wide margin of discretion when defining and moderating online content. This powerful role played by OSPs is further enhanced by the confidentiality that reigns in the content moderation realm. Consequently, users may not always understand what is (im)permissible content when they use an online platform, or why and how a moderation decision has been taken in relation to the content they posted.

Even if there is little transparency with respect to the online content moderation process, research confirms that technical (often AI) tools and human moderation nowadays go hand in hand to combat cyberviolence. Technical tools seem to be used primarily to prevent (i.e., proactively) or detect (i.e., reactively) impermissible content, while human moderators rarely intervene proactively. Instead, they react to user notifications or to content flagged by AI tools. Moreover, more recently established online platforms turn more often to (human) non-professional content moderators.

The Digital Services Act of the EU will considerably impact the role of service providers in combating illegal online content. It imposes a whole range of new due diligence obligations on OSPs as well as an obligation to cooperate with LEAs. These obligations vary depending on the service and size of the provider, but will lead to more transparency on how providers deal with impermissible content.

Coping mechanisms and victim support

The study on emotions and coping mechanisms of victims of OHS and NCII shows the substantial impact of these behaviours on adolescents and young adults. Among the respondents, the most common feelings after victimisation of OHS are feeling angry (46.9%) or furious (41.4%). Further, one third of the victims of OHS feels nervous (36.8%), ashamed (32.2%), helpless (32.6%) and irritated (37.7%), whilst one out five victims feels guilty (21.3%) and lonely (22.6%). For NCII, the most reported feelings associated with victimisation are nervousness (52.6%), helplessness (55.3%), anger (48.4%), guilt (47.4%), and irritation (48.8%). Further, more than 40% of NCII victims feel anxious (43.2%), ashamed (46.3%), lonely (44.2%) or furious (42.6%). The fact that both for OHS and particularly for NCII high prevalence of feelings of loneliness, nervousness and helplessness are experienced, indicate the psychological impact these behaviours may have on victims.

Notwithstanding the substantial impact and harm of victimisation of OHS and NCII, the study shows that adolescents and young adults only scarcely reach out for professional help, including police or victim support organisations. In turn, victim support organisations indicate that adolescents and young adults will not easily reach out to them. Rather, adolescents and young adults will discuss their experiences with their peers and to a lesser extent with their relatives (e.g., parents). As such, it is advisable to improve the knowledge of adolescents and young adults on the potential of support organisations in coping, to remove potential hurdles by a low-threshold access and to invest in wider communication to the wider public, particularly to young people in school and parents, on coping and support. Given that peers are often the first contact for victims, it is worth to invest in active bystanding programmes for adolescents and young adults, where they learn how to support and inform victims.

The study further shows that there is a vast network of organisations in Belgium that provide support for victims of cyberviolence, either based on the behaviour (OHS or NCII) or on characteristics of the victims (gender or sexual orientation). There appears to be a strong informal cooperation between the networks. However, due the lack of a formal coordination and structure as well as due to the budgeting to address cyberviolence often happens on a project-based ground, there is overlap of energy and resources. All organisations highlighted the importance of mainstreaming information on boundaries online in formal education.

An important step to further improve the efficiency and role of the support organisations is to ensure that they have sufficient budget and capacity to live up to the requirements of the DSA with regard to acting as trusted flaggers, i.e., hotlines with a prioritised connection to OSPs. This would mean that, when these organisations flag unlawful OHS or NCII, OSPs will be required to act quickly upon their reports. Currently, three organisations already informally function as 'trusted flaggers' but report an incoherent and uneven cooperation of the OSPs. This could change once they acquire the status of DSA trusted flagger. However, in turn they will have to ensure qualitative and efficient hotlines, which will only be possible with the necessary investments in technology and capacity.

Overall conclusions

In view of the holistic analysis of the study results, the @ntidote team decided on ten main findings that further shape the understanding of OHS and NCII as well as the current approach to these behaviours:

1. There is no common understanding of what constitutes cyberviolence, including what constitutes OHS and NCII. This complicates research as well as prevention.
2. Encounters with OHS and NCII are highly prevalent among adolescents and young adults. In most cases they are bystanders, but there is also a significant group that is victimised.
3. Contrary to common perception, there is a wide variety of motives associated with perpetration of OHS and NCII.
4. Relevant criteria for victimisation and perpetration for both OHS and NCII are age and ethnicity. Sexual orientation is a significant criterion for victimisation of OHS. Gender was found to be a relevant criterion for perpetration of NCII.
5. Notwithstanding a developed legal framework that criminalises (forms of) OHS and NCII, the vast majority of criminal complaints are discontinued. The lack of capacity and prioritisation are recurrent reasons for the high level of discontinued cases.

6. Specifically for OHS, the procedural hurdle for prosecuting cases before the Court of Assize results in the discontinuation of many cases and is considered problematic, both from the perspective of the European and international legal framework as from the perspective of support organisations.
7. Adolescents and young adults are significantly in favour of alternatives to classic criminal punishments when addressing criminal complaints on OHS and NCII.
8. A limited number of major OSPs are predominantly used by adolescents and young adults. Certain OSPs are prevalent in the criminal reports, interviews, and prevalence study in relation to occurrence of OHS and NCII.
9. Whereas OSPs are considered vital in tackling OHS and NCII, analysis shows that there is a wide variety among the OSPs in the delineation of permissible and non-permissible content and their procedures of moderation and removal. Support organisations highlight that the cooperation with OSPs for the removal of OHS or NCII is incoherent.
10. Victims of OHS and NCII experience substantial harm and negative emotions, but generally do not reach out to professionals, including support organisations or police. They will mostly turn to peers for support.

4.2. RECOMMENDATIONS

Based on the research, the @ntidote study has drafted several recommendations. Regarding several of these recommendations, (members of) the @ntidote team has already taken steps or sought collaboration (see 5. Dissemination and valorisation). These recommendations were further categorised in the following themes:

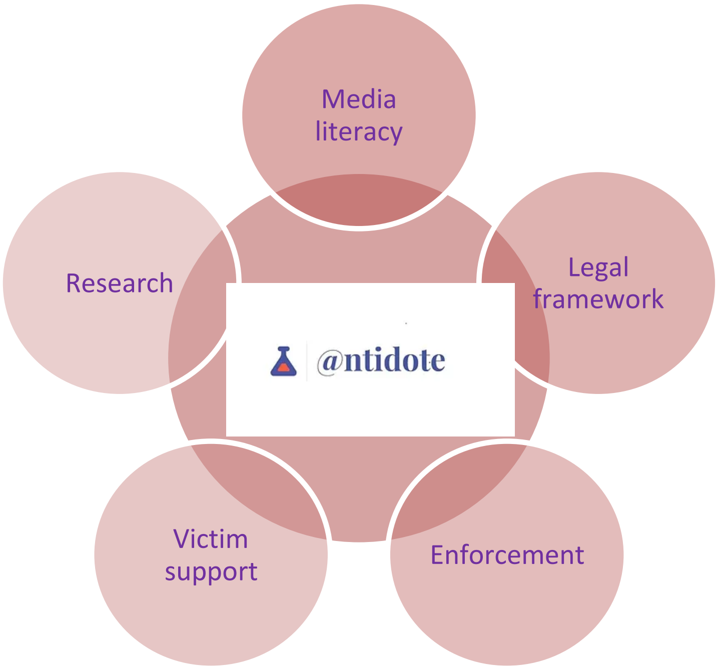


Figure 30. @ntidote recommendations

4.2.1. MEDIA LITERACY

R1. Improve the understanding among both adolescents and young adults as well as the common public on what constitutes OHS and NCII, including as to the delineation of unlawful speech.

R2. Invest in the development of societal tolerance principles and changing attitudes on OHS towards minority groups with a particular focus on discussing with adolescents and young adults the non-acceptability of OHS.

R3. Stimulate education and discussion in schools on online boundaries, understanding of what constitutes OHS and NCII and the impact of these behaviours to change attitudes. Train teachers in media literacy, including on OHS and NCII.

R4. Include adolescents and emerging adults - spanning the diversity of gender, age, sexual orientation, and ethnicity – in developing solutions for OHS and NCII via co-creation.

R5. Raise awareness within the general population on the harmfulness of OHS and NCII and address victim-blaming.

R6. Support active bystandership online so that witnesses of cyberviolence can act when confronted with NCII or hate speech, to decrease perpetration and support the victim. Improve in this regard knowledge on effective bystanding.

R7. Collaborate with organisations that are specialised in working with minority groups to mainstream knowledge on the harmfulness of OHS and NCII as well as what actions to take in case of victimisation.

4.2.2. LEGAL FRAMEWORK

R8. Reconsider the current national legislation and procedural hurdles for prosecution of OHS in the light of the international and European supranational legal framework on OHS.

R9. Improve and monitor alternatives to prosecution before courts, such as mediation and probation trajectories aiming for restoration and behavioural change.

R10. Reopen a national dialogue to reconsider which forms of hate speech are to be criminalised as well as alternatives to criminalisation to tackle OHS within the boundaries of the international and European legal framework.

R11. Consider acceding to the first additional protocol of the Budapest Cybercrime Convention that explicitly addresses criminalisation and cooperation in cases of OHS.

R12. Support the development at the EU level of what constitutes unlawful content, in particular regarding OHS and NCII, in order to create a common denominator for removal on all OSPs active in the EU.

R13. Support the development of international and supranational norms on NCII based on the principles of equality, the prohibition of gender-based violence, and sexual integrity, whereby consent should be the defining element for the delineation of unlawful dissemination of intimate images.

4.2.3. ENFORCEMENT

R14. Improve skills and appreciation of law enforcement on OHS and NCII via training and guidelines and streamline grounds of prosecution and discontinuation for OHS and NCII.

R15. Invest in capacity of specialised police to investigate and act against OHS and NCII.

R16. Clarify the categorisation of OHS and NCII in databases of the police, the public prosecutor's office, and courts, to have a better overview and enable future analysis of the case law.

R17. Provide guidelines to prosecution on the delineation between NCII and CSAM in order to improve best fit qualification and prosecution.

R18. Develop and apply alternatives to classic punishments, such as prison sentences, for perpetration of OHS and NCII. Develop in this respect a compulsory course specific for perpetrators of OHS and NCII within the framework of probation or mediation.

R19. Closely monitor and enforce the implementation of the DSA, both at the EU and at the national level, especially regarding the new due diligence and cooperation obligations.
R20. Invest in discussions as well as enforcement of the DSA, particularly in relation to those OSPs that are prevalent in occurrence of NCII and OHS victimising adolescents and emerging adults.

4.2.4. VICTIM SUPPORT

R21. Ensure that psychological help for victims of cyberviolence is accessible and available. Communicate the relevance of psychological support.
R22. Improve knowledge among adolescents and emerging adults, parents, and schools how to support victims and where to find professional help.
R23. Convince media to include the contacts of support organisations when publishing articles on OHS or NCII.
R24. Convince OSPs to publish a list of trusted flaggers on the website to redirect victims and bystanders to national expertise victim support organisations.
R25. Invest in capacity for and coordination among victim support organisations for OHS and NCII via structural budgets for their roles as victim support and trusted flaggers.
R26. Stimulate specialised victim support organisations to acquire the status of trusted flagger under the EU Digital Services Act.
R27. Decrease hurdles for contacting victim support organisations, e.g., by investing in outreach as well as technology that allows for a first anonymous contact (such as chatboxes).
R28. Incentivise the development of technology and cooperation between OSPs, authorities, and support organisations to prevent, find, and remove unlawful content.

4.2.5. RESEARCH

R29. Develop a vocabulary on cyberviolent behaviours that can be understood and used by digital natives.
R30. Gain a better understanding of how peer communication and regulation work in online discussion groups, especially in the context of NCII.
R31. Include adolescents and emerging adults - spanning the diversity of gender, age, sexual orientation, and ethnicity – in research to gain further insight in certain dynamics and their understanding of behaviours.
R32. Develop the human rights framework delineating NCII in the light of artificial forms of intimate images in the light of denuding technology, particularly concerning the freedom of expression, right to information and freedom of press.
R33. Invest in further research on both the underpinning of differentiations in emotions and coping mechanisms of victims of OHS and NCII to better understand support needs.
R34. Further research in-depth the role of personal characteristics of adolescents and emerging adults by conducting research that includes minorities only.
R35. Support qualitative research on OHS and NCII in the intersectional population, namely in adolescents and emerging adults who were both victim and perpetrator.
R36. Collaborate with organisations that are specialised in working with minority groups to access bigger samples for research purposes.
R37. Conduct further research on proactive and reactive content moderation, especially with respect to trusted flaggers, the follow-up given to user notifications, the remedies available to users, and the treatment of impermissible content.
R38. Stimulate further research on the content moderation process to better understand the factors relevant for defining the permissibility of NCII and OHS, the role of consent, age, ethnicity, and other personal characteristics.

5. DISSEMINATION AND VALORISATION

5.1. Project Website and social networks

- Website: https://www.antidoteproject.be/
- Facebook:https://www.facebook.com/people/Antidote-Project Belspo/100078679002260/?ref=py_c
- X/Twitter: https://x.com/antidote4cyber1?s=20

5.2. International expert seminar on evidence-based cyberviolence policy in Europe

An international research seminar on evidence-based cyberviolence policy in Europe – 8-9 December 2022, University of Antwerp was organised by the @ntidote team (8-9 December 2022, University of Antwerp) and its three researchers gave a presentation at this seminar:

- Gangi, O., "Online hate speech among digital natives: definitions, perceived motivations and feelings of harm by sexual orientation", 09/12/2022.
- Giacometti, M., "Non-consensual distribution of intimate images of adult and minor victims: two offences and some overlaps", 08/12/2022.
- Gilen, A., "The roles of gender and sexual orientation in mapping the psychosocial harm and coping strategies in NCII victims", 08/12/2022.

5.3. Conferences and seminars

- Eurocrim 2023: 23rd Annual Conference of the European Society of Criminology (Florence) - intervention by Catherine Van de Heyning, Michel Walrave, Mona Giacometti, Aurélie Gilen & Amber Van de Maele, on the topic *"The non-consensual possession of intimate images in adolescents and emerging adults",* 07/09/2023.
- Eurocrim 2023: 23rd Annual Conference of the European Society of Criminology (Florence) - poster prepared by Océane Gangi and Cécile Mathys on the topic "*Why do we share intimate images of others? Perceptions of 15 to 25 years old Belgian youths",* 07/09/2023.
- Eurocrim 2023: 23rd Annual Conference of the European Society of Criminology (Florence) – organisation of a panel by Michel Walrave entitled "Image-based sexual abuse: Observations and new trends" with the following presentations: *"Exploring risky online sexual behaviour amongst European Youth: Findings from an H2020 study".* Julia C. Davidson (University of East London, UK), Mary Aiken (Capital Technology University, USA, University of East London, UK), Kirsty Phillips (University of East London, UK), Ruby Farr, University of East London, UK; "*Disrupting and preventing sexualised deepfake abuse: Findings from a multi-country study".* Asher Flynn (Monash University, AUS) , Anastasia Powell (RMIT University, AUS) , Adrian J. Scott (University of Goldsmiths, UK), Elena Cama (University of New South Wales, AUS); *"Sext dissemination: a systematic review and research agenda"* Silke Van den Eynde (KU Leuven), Stefaan Pleysier (KU Leuven), Michel Walrave; "Different manifestations of Image-Based Sexual Abuse within Telegram Groups" Edel Beckman (PermessoNegato, IT), *Cosimo Sidoti* (Università Cattolica del Sacro Cuore and Transcrime, IT)
- International conference organised by the University of Luxembourg on Private actors as judges and enforcers in the technology-driven world (Luxembourg), intervention by Vanessa Franssen, on the topic *"Online content moderation: the invisible hand of intermediary service providers in the fight against cyberviolence",* 04/07/2023.
- Conference organised by Université Libre de Bruxelles (Brussels), intervention by Mona Giacometti, on the topic *"@ntidote: toward a better understanding of cyberviolence in Belgian legal practice"*, 01/06/2023.

- Seminar organised by ERA (Academy of European Law) (webinar), intervention by Catherine Van de Heyning and Mona Giacometti, on the topic *"Online Sexual Violence and Image-based Abuse: Investigation, Prosecution and Litigation",* 10/05/2023.
- Conference organised by USL-B and UCLouvain on The implementation of the Digital Services Act: Responsibilities, new due diligence obligations and enforcement issues (Brussels), intervention by Vanessa Franssen and Marine Corhay, on the topic *"Illegal content: The case of hate speech",* 03/05/2023.
- Seminar by Jura Falconis on Sexual Boundary Crossing Behavior and the New Sexual Criminal Law (Leuven), intervention by Mona Giacometti, on the topic *"Digitaal seksueel grensoverschrijdend gedrag"*, 24/03/2023.
- Technology Law Session (Antwerp), intervention by Catherine Van de Heyning and Mona Giacometti on the topic *"Online hate speech",* 01/02/2023.
- Festival Van de Gelijkheid (Ghent), intervention by Catherine Van de Heyning on the topic "Wat kan je doen tegen wraakporno?", 16/12/2022.
- International Conference on (Cyber)bullying Critical and interdisciplinary approaches of online violence phenomena (Nancy), intervention by Océane Gangi, on the topic *"Interrelations entre le discours de haine en ligne et le cyberharcèlement: quelles spécificités retrouve-t-on au sein d'un échantillon de digital natives?"*, 07/12/2022.
- Commission Université-Palais (Louvain-la-Neuve & Charleroi), intervention by Mona Giacometti, on the topic *« Les discours de haine en ligne : vers un cadre légal plus moderne ? »*, 18/11/2022 & 02/12/2022.
- Human Factor in Cybercrime (Florida), intervention by Aurélie Gilen, on the topic "*The non-consensual dissemination of intimate images (NCII): victims' rationale behind not reporting this crime and their perspective on how to legally conserve NCII*", 22/11/2022.
- Dépasser les bornes, organised by University of Liege (Liège), intervention by Mona Giacometti, Océane Gangi and Aurélie Gilen, on the topic *« Diffusion non consentie de contenus à caractère sexuel et diffusion d'images d'abus sexuels de mineurs : entre distinctions et chevauchements, quelles implications d'un point de vue légal, criminologique et psycho-social? »*, 07/10/2022.
- Eurocrim 2022: 22rd Annual Conference of the European Society of Criminology (Malaga), intervention by Catherine Van de Heyning, Aurélie Gilen and Michel Walrave, participation of Mona Giacometti, on the topic *"Coping with online sexual image-based abuse: strategies of victims and bystanders"*, 22/09/2022.
- Eurocrim 2022: 22rd Annual Conference of the European Society of Criminology (Malaga), poster by Aurélie Gilen, Océane Gangi, Catherine Van de Heyning, Cécile Mathys and Michel Walrave, on the topic "*Non-consensual dissemination of intimate images: do victims' and bystanders' perspectives align with the current sanctions recorded in the criminal code?*", 22/09/2022.
- Hanna Arendt Institute (online), intervention by Catherine Van de Heyning on the topic "online hate speech", 21/06/2022.
- « Le nouveau droit pénal sexuel », colloquium organised by the conference of the Young Bar of Brussels (Brussels), intervention by Mona Giacometti on the topic "*Voyeurisme et diffusion non consentie d'images à caractère sexuel. Maintien du statu quo ou réelles nouveautés ?",* 02/06/2022.
- UNIA, Centre for equal opportunities and opposition to racism, contribution by Michel Walrave and Catherine Van de Heyning in a report on the topic "*Policy input to the online sexual violence policy plan of the State Secretary for Gender Equality*", spring 2022.
- « Criminal justice and digitalization », International research seminar organised by Vanessa Franssen (Liège), intervention by Mona Giacometti on the topic *"@ntidote: Toward a better understanding of cyberviolence in Belgian legal practice*", 16/05/22.

- Colloquium organised by AICLF (Association internationale des criminologues de langue française) (Ottawa), intervention by Océane Gangi and Cécile Mathys on the topic: *« Entre discours de haine en ligne et cyberharcèlement chez un public de 15 à 25 ans : une distinction de fait et de droit, mais une distinction pertinente en criminologie ? »,* 15/05/22.
- Commission Justice of the Federal Parliament (Brussels), intervention by Catherine Van de Heyning on the topic "Online sexual violence and the sexual criminal law", 19/10/2021.
- NVKVV conference (Network of nurses, conference during the nurses' week) (Ostend), intervention by Michel Walrave, on the topic "*Cyber violence: types, impact, prevention and intervention"*, 29/09/2021.
- University of Amsterdam, intervention by Jogchum Vrielink on the topic "debate on lawsuits for hate speech against politicians", 21/09/2021

5.4. Forthcoming presentations

CIFAS 2024 (Lausanne), intervention by Océane Gangi and Cécile Mathys, on the topic "*On a tous un dossier de nudes sur son téléphone" : A la rencontre des expériences subjectives de partage non consenti d'images intimes de jeunes belges âgés de 15 à 25 ans*, xx/06/2024.

5.5. Policy Briefs

Hate speech among Belgian youth aged 15 to 25: "50 Shades of Hate Speech" (Belspo Magazine).
On the legal aspects of online hate speech: "Free the bird" (Belspo Magazine).

5.6. Others

- Nieuwsblad and Gazet Van Antwerpen, intervention by Michel Walrave, on the topic "Wat bezielt jongeren die bruut pestgedrag filmen en delen? "Dit niveau van geweld en intensiteit baart mij zorgen", "Wat bezielt jongeren die bruut pestgedrag filmen én delen? De psychologie achter de 'happy slappers', 20/08/2023. https://www.nieuwsblad.be/cnt/dmf20230819_97576338; https://www.gva.be/cnt/dmf20230819_97666346
- Het Laatste Nieuws, intervention by Catherine Van de Heyning and Michel Walrave, on the topic Tieners slaan andere tieners in elkaar én filmen dat. Experts geven tips voor ouders: "De 5 A's zijn de meest effectieve methodes", 22/06/2023 https://www.hln.be/binnenland/tieners-slaan-andere-tieners-in-elkaar-en-filmen-dat-experts-geven-tips-voor-ouders-de-5-as-zijn-de-meest-effectieve-methodes~a07e264e/
- Het Laatste Nieuws, intervention by Catherine Van de Heyning and Michel Walrave, on the topic "Kinderen wisselen naaktbeelden van zichzelf uit zoals Pokémonkaarten": politie raadt ouders aan om gesprek aan te gaan", 27/07/2023 https://www.hln.be/binnenland/kinderen-wisselen-naaktbeelden-van-zichzelf-uit-zoals-pokemonkaarten-politie-raadt-ouders-aan-om-gesprek-aan-te-gaan-br~aa2b4d62/
- Knack, intervention by Catherine Van de Heyning and Michel Walrave, on the topic "Wat wil je dat ik voor je doe? Kinderen maken vakker hun eigen misbruikbeelden. 26/07/2023 https://www.knack.be/nieuws/belgie/maatschappij/wat-wil-je-dat-ik-voor-je-doe-kinderen-maken-vaker-hun-eigen-misbruikbeelden/
- Knack, intervention by Catherine Van de Heyning on the topic "Cybercriminaliteit: *'Voor jongens zijn naaktbeelden als Pokémonkaarten"*, 13/07/2023.
- VRT NWS on Youtube, intervention by Aurélie Gilen on the topic "*EDUbox Sexting: Artificiële Intelligentie"*, 31/05/2023.
- Workshop for students, intervention by Océane Gangi, on the topic "*Toi, quelles sont tes limites dans ton couple ?"*, 15/02/2023.

- Nieuwsblad, intervention by Catherine Van de Heyning, on the topic "*Ongewenste dickpics sturen wordt strafbaar: 'Ze staan er niet eens bij stil hoe agressief zo'n foto kan zijn'*", 24/11/2022. https://www.nieuwsblad.be/cnt/dmf20221124_96693075
- De Morgen - intervention by Catherine Van de Heyning, on the topic *"'Naaktfoto's en info van meisjes massaal gedeeld in 'exposegroepen': 'De meesten weten zelf niet dat het gebeurt",* 15/11/2022. https://www.demorgen.be/tech-wetenschap/naaktfoto-s-en-info-van-meisjes-massaal-gedeeld-in-exposegroepen-de-meesten-weten-zelf-niet-dat-het-gebeurt~b431547a/
- Studio Brussel - Faqda, intervention by Catherine Van de Heyning and Aurélie Gilen, on the topic *"Haat: Zit haat in ons?",* 15/10/2022. https://www.vrt.be/vrtmax/a-z/faqda/5/faqda-s5a7/
- Expert panel about the film #salepute (Ghent), intervention by Aurélie Gilen, on the topic *"Legal, sociological, and psychological dimensions of cyberviolence against women"*, 12/10/2022.
- De Morgen - intervention by Catherine Van de Heyning, on the topic *"Doe alleen aan sexting op apps die daarvoor dienen': experte waarschuwt voor online afpersing"*, 19/08/2022. www.demorgen.be/nieuws/doe-alleen-aan-sexting-op-apps-die-daarvoor-dienen-experte-waarschuwt-voor-online-afpersing~b8b3f411/
- VRT NWS, intervention by Catherine Van de Heyning on the topic *"Online afpersing via Snapchat en Tinder, verkrachting en aanranding: grote zedenzaak uitgesteld naar oktober",* 18/08/2022. https://www.vrt.be/vrtnws/nl/2022/08/17/online-afpersing/
- Het Laatste Nieuws, intervention by Michel Walrave and Catherine Van de Heyning, on the topic *"Ouders spelen een grote rol":* hoe bescherm je jezelf tegen 'sextortion'? Experts geven tips", 18/05/2023. www.hln.be/binnenland/ouders-spelen-een-grote-rol-hoe-bescherm-je-jezelf-tegen-sextortion-experts-geven-tips~a98d0b02/
- De Standaard, intervention by Catherine Van de Heyning, on the topic *"Het is alsof de verkrachting jarenlang online voortging",* 07/05/2022. https://www.standaard.be/cnt/dmf20220505_96186126
- De Standaard, intervention by Catherine Van de Heyning, on the topic *"Voor daders is er een groot verschil tussen online en fysiek seksueel geweld, voor slachtoffers niet",* 10/02/2022. https://www.standaard.be/cnt/dmf20220505_96186126
- De Standaard, podcast of Catherine Van de Heyning, on the topic *"Voor daders is er een groot verschil tussen online en fysiek seksueel geweld, voor slachtoffers niet",* 10/02/2022. https://www.standaard.be/cnt/dmf20220210_93667473
- VRT Radio 1, intervention by Catherine van de Heyning on the topic *"Cyberpesten is een fenomeen dat we steeds meer zien oprukken door sociale media"*, 01/02/2022.

5.7. Forthcoming

The team will organise an event on 17 November 2023 at the University of Saint-Louis, bringing together 3 schools from the 3 regions (Flanders, Brussels, Wallonia). The event will include roundtables to discuss NCII and OHS, and will be a unique opportunity to communicate our results to the media and to share scientific output with the target population of our research project.

6. PUBLICATIONS

6.1. Previous publications

Peer-reviewed

- Gangi, O., Brassine, N. & Mathys, C. (2023). "Entre discours de haine en ligne et cyberharcèlement chez un public belge de 15 à 25 ans". *Criminologie, Forensique, et Sécurité*, 1 (1), 3620.
- Gangi O., Giacometti M. & Gilen A., (2022). "Diffusion non consentie de contenus à caractère sexuel et diffusion d'images d'abus sexuels de mineurs : entre distinctions et chevauchements, quelles implications d'un point de vue légal, criminologique et psycho-social *?", Revue de la Faculté de Droit de l'Université de Liège*, (3), 635-374.
- Giacometti M. (2022). "Les discours de haine en ligne : vers un cadre légal plus moderne?", in V. Franssen & A. Masset (eds), *Le droit pénal et la procédure pénale en constante évolution*, Commission Université-Palais, Anthemis, 171-203.
- Giacometti, M. (2022) "Voyeurisme et diffusion non consentie d'images à caractère sexuel. Maintien du statut quo ou réelles nouveautés ?", in *Le nouveau droit pénal sexuel*, Bruxelles, Larcier, 143-186.
- Gilen, A. & Vreven, N. (2022). De digitale dimensie van seksuele zelfexpressie: een bevrijding of een nieuwe weg naar criminaliteit? *Tijdschrift van Mensenrechten*, 4-10.
- Lemmens, K., & Vrielink, J., (2022). "Nazinderende geschiedenis. Enkele bedenkingen bij de bestraffing van het gebruik van nazisymbolen in het licht van de vrijheid van meningsuiting", in *De Grondwet en Jan Velaers*, Bruges, die Keure, 115-222.
- Van de Heyning, C. & Giacometti, M. (2023). Haatspraak of hatelijke belaging: tijd voor een nieuwe kijk op artikel 150 van de Grondwet. *Tijdschrift voor Mensenrechten*, 21(1), 6-13.
- Van de Heyning, C. & Giacometti, M. (2022). Het verspreiden van naaktbeelden zonder toestemming krijgt verder uitwerking op nationaal niveau. *Tijdschrift voor Strafrecht*, 157-160.
- Van de Heyning, C. & Giacometti, M. (2023). "Online seksueel geweld: daderschap herbekeken", in C. Mussche & L. Stevens (eds). *Onderzoek en preventie van seksuele misdrijven.* Intersentia.
- Vrielink, J. (2022). « Bedenkingen inzake voorstel tot herziening van artikel 25 van de Grondwet en voorstel tot herziening van artikel 160 van de Grondwet », *Cahiers de CIRC*, (5), 52-65.
- Walrave, M. & Van de Heyning, C. (2022). "De beelden waren de druppel: waarom beelden van seksuele misdrijven online gedeeld worden vanuit sociaalwetenschappelijk en juridisch perspectief". *Cahier Politiestudies*, 62(1), 163-189.

Others

- Van de Heyning, C. (2021). Hoe wenselijk zijn anticiperende social media? Samenleving & Politiek 28:3, 57-62.
- Van de Heyning, C. (2021). Volodina t. Rusland nr 2 - Cybergeweld in het vizier van Straatsburg. ECHR Updates 14 September 2011.

6.2. Forthcoming publications

- Franssen, V., Giacometti, M., & Corhay, M., "The Digital services Act and the fight against cyberviolence: New rules regulating the liability of OSPs for illegal content, stronger protection for users?", *Common Market Law Review*.
- Franssen, V., Giacometti, M. & Van de Heyning, C., "Social media perspectives of the role of trusted flaggers tackling cyberviolence", *in Cyberviolence: towards an evidence-based policy on online harm in Europe,* Edward Elgar publishing.
- Franssen, V., Gangi, O., Giacometti, M. & Gilen, A., "Online content moderation through the eyes of moderators: A revealing look inside the black box", to be published in an American journal (e.g. Harvard Journal of Law & Technology, Yale Journal of Law and Technology, Georgetown Law Technology Review or Fordham Law Review).

- Gangi O, & Mathys, C. (in press). “Discours de haine en ligne : vers une plus grande compréhension des définitions et des expériences d’un public cible âgé de 15 à 25 ans en Belgique selon les caractéristiques individuelles”, in *Ouvrage collectif CIC,* Éditions de l'Université de Lorraine, France.
- Gangi, O., Giacometti, M. & Gilen, A., **“**Non-consensual dissemination of intimate images of adult and minor victims: two offences and some overlaps”, *in Cyberviolence: towards an evidence-based policy on online harm in Europe,* Edward Elgar publishing.
- Gangi, O., & Mathys, C., “The perceived motivations behind online hate speech and non-consensual dissemination of intimate images”, *Victims & Offenders*.
- Gangi, O., & Mathys, C., “Online Hate Speech among Belgian Digital Natives: Focus on Gender and Sexual Orientation”, in *Cyberviolence: towards an evidence-based policy on online harm in Europe*, Edward Elgar publishing.
- Giacometti, M. & Van de Heyning, C., “Le rôle des fournisseurs de services en ligne dans la lutte contre les cyberviolences : le cas de la diffusion non consentie d’images intimes”, *Conférence du jeune barreau de Bruxelles,* Bruxelles, Larcier.
- Giacometti, M., Franssen, V. & Claes, A.L., “OSPs’ definitions of cyberviolence: The case of online hate speech and non-consensual dissemination of intimate images”, *European Journal of Crime, Criminal Law and Criminal Justice*.
- Gilen, A. & Walrave, M. “*The roles of gender and sexual orientation in mapping the psychosocial harm and coping strategies for victims of non-consensual dissemination of intimate images”,* in *Cyberviolence: towards an evidence-based policy on online harm in Europe,* Edward Elgar publishing.
- Van de Heyning, C., Keiler, J. & Franssen, V., “De strafbaarstelling van digitaal seksueel beeldmisbruik in de Lage Landen”, *Cahier Politiestudies*, 70, February 2024.
- Vrielink, J. & Lemmens, K. “Hate speech, het EHRM en de Belgische rechtspraak: een ongeliefde inperking van de uitingsvrijheid”, in S. Rutten e.a. (eds.), *Recht en diversiteit, Cambridge/Antwerpen, Intersentia*, 2023, 125-156
- Van de Heyning, C. & Walrave, M. “Online seksueel geweld: Daderschap anders bekeken” in C. Mussche and L. Stevens (eds.) *Onderzoek en preventie van seksuele misdrijven*, *Cambridge/Antwerpen, Intersentia*, 2023
- Walrave, M., Van de Heyning, C., Janssen, J., & Kolthoff E. (eds) “Politie en misbruik van beelden”. *Cahiers Politiestudies*,70, February 2024.
- Walrave, M., Schokkenbroek, J., Gilen, A., Ponnet, K. & Hardyns ,W. “Digitaal partnergeweld: Typology, impact en rol van politie”. *Cahiers Politiestudies*, 70, February 2024.

6.3. Forthcoming book

Van de Heyning, C., Walrave, M., Franssen, V., Mathys, C. & Vrielink, J. (eds), *Cyberviolence : towards an evidence-based policy on online harm in Europe*, Edward Elgar publishing.

7. ACKNOWLEDGEMENTS

Regular news coverage shows that investigating, preventing, and addressing online hate speech and non-consensual dissemination of intimate images remains undoubtedly an important societal challenge. The issues addressed in the @ntidote project are, therefore, relevant and urgent. This two-year project has been a strong collaboration between researchers from different disciplines that resulted in the project's scientific output but also numerous activities of public outreach. This collaboration also led to new research projects and other initiatives.

We are, therefore, first of all grateful for the opportunity offered by BELSPO to conduct this interdisciplinary research in a domain that is topical for national and international policymakers and other stakeholders working on preventing and countering hate speech and non-consensual dissemination of intimate images. During the project, we had valuable input from the members of our follow-up committee and would, therefore, like to thank them for their support and inspiration. We hope to further collaborate with them in the framework of our follow-up research in this domain.

We would like to thank the brilliant students who helped us out during the project, in particular with the coding of the many judgments and criminal complaints, namely Loris Rossi, Victoria Gilles, Noa Vreven, Oliver Wouters and Charlotte Dierickx-Visschers. In the final stages we were also aided by two additional researchers, Marine Corhay and Ana Laura Claes, to whom we extend our gratitude.

The @ntidote team is also grateful for the time and effort experts from several organisations invested in this project, by participating in interviews and sharing their experience with us. More particularly, we would like to thank çavaria, Child Focus, Mediawijs, the Institute for Equality of Women and Men, Sensoa, Unia, and experts from several OSPs and industry associations that collaborated throughout the project. The sharing of their day-to-day experience in the field was a great inspiration and support.

Further, the legal analysis would not have been possible without the support of the prosecution offices of Antwerp, East Flanders, Halle-Vilvoorde, Brussels, and Namur as well as the courts of Liège, Brussels, and Antwerp. We are particularly grateful to the administrative support of the prosecution offices and the courts for helping us navigate the many files.

Finally, we would particularly like to thank those adolescents and young adults who participated in the interviews and the survey. They are at the core of this project. By sharing their often very intimate experiences, we could learn and develop solutions to help their generation and coming ones.

The @ntidote team

8. REFERENCES

Al Serhan, F. et Elareshi, M. (2019). University Students' Awareness of Social Media Use and Hate Speech in Jordan. *International Journal of Cyber Criminology*, *13*(2), 548–563. https://doi.org/10.5281/zenodo.3709236

Alonso, C., & Romero, E. (2019). Conducta de sexting en adolescentes: predictores de personalidad y consecuencias psicosociales en un año de seguimiento. *Anales de psicología / Annals of Psychology, 35*(2), 214-244. https://doi.org/10.6018/analesps.35.2.339831

Anti-Defamation League (2018). *The Pyramid of hate*. Retrieved from https://www.adl.org/sites/default/files/documents/pyramid-of-hate.pdf .

Aranda Juárez, D., Sánchez-Navarro, J., & Mohammadi, L. (2020). Perception and self-assessment of digital skills and gaming among youth: A dataset from Spain. *Data in Brief, 28*, 104957–104957. https://doi.org/10.1016/j.dib.2019.104957

Arsht, A., & Etcovitch, D. (2018, March 2), "The Human Cost of Online Content Moderation", *Harvard Journal of Law & technology*.

Aswad, E., (2018) The Future of Freedom of Expression Online. *Duke Law & Technology Review 26*, Available at SSRN: https://ssrn.com/abstract=3250950

Awan, I. (2014). Islamophobia and Twitter: A Typology of Online Hate Against Muslims on Social Media: Islamophobia and Twitter. *Policy and Internet, 6*(2), 133–150. https://doi.org/10.1002/1944-2866.POI364

Baider, F. (2019). Le discours de haine dissimulée : le mépris pour humilier. *Déviance et société*, 43(3), 359–387. https://doi.org/10.3917/ds.433.0359

Bal, S., Crombez, G., De Bourdeaudhuij, I., & Van Oost, P. (2009). Symptomatology in adolescents following initial disclosure of sexual abuse: The roles of crisis support, appraisals and coping. *Child Abuse & Neglect*, *33*(10), 717-727. https://doi.org/10.1016/j.chiabu.2008.11.006

Bandura, A. (1976). Self-Reinforcement: Theoretical and Methodological Considerations. *Behaviorism, 4*(2), 135–155.

Barrett, P. (2020) Who moderates the social media giants? A call to end outsourcing. NYU Stern Center Centre for Business and Human Rights. https://www.stern.nyu.edu/experience-stern/faculty-research/who-moderates-social-media-giants-call-end-outsourcing

Bates, S. (2017). Revenge porn and mental health: a qualitative analysis of the mental health effects of revenge porn on female survivors, *Feminist Criminology 12*(1), 22–42

Bautista-Ortuño, R., Perea García, J., Rodríguez Gómez, N. & Castro Toledo, F. (2018). "May I offend you?" An experimental study on perceived offensiveness in online violent communication and hate speech. *International e-journal of criminal sciences*, 12.

Beausoleil, L (2019). Free, Hateful, and Posted: Rethinking First Amendment Protection of Hate Speech in a Social Media World, *Boston College Law Review 60(*7), 2101 – 2144

Bedrosova, M., Machackova, H., Šerek, J., Smahel, D., & Blaya, C. (2022). The relation between the cyberhate and cyberbullying experiences of adolescents in the Czech Republic, Poland, and Slovakia. *Computers in Human Behavior*, *126*, 107013. https://doi.org/10.1016/j.chb.2021.107013

Beliveau, A (2018). Hate Speech Laws in the United States and the Council of Europe: The Fine Balance between Protecting Individual Freedom of Expression Rights and Preventing the Rise of Extremism and Radicalization through Social Media Sites Notes, *Suffolk University Law Review 51*(4), 565 – 588

Bellanova, R., & de Goede, M. (2022). Co-Producing Security: Platform Content Moderation and European Security Integration. *Journal of common market studies*, *60*(5), 1316–1334. https://doi.org/10.1111/jcms.13306

Bennett, S., Maton, K., & Kervin, L. (2008). The 'digital natives' debate: A critical review of the evidence. *British Journal of Educational Technology*, *39*(5), 775786. https://doi.org/10.1111/j.1467-8535.2007.00793.x

Bernatzky, C., Costello, M., & Hawdon, J. (2022). Who Produces Online Hate? An Examination of the Effects of Self-Control, Social Structure, & Social Learning. *American Journal of Criminal Justice, 47*(3), 421–440. https://doi.org/10.1007/s12103-020-09597-3

Beyens, J. & Lievens, E. (2016) A legal perspective on the non-consensual dissemination of sexual images: Identifying strengths and weaknesses of legislation in the US, UK and Belgium, *International Journal of Law, Crime and Justice 47* (December issue), 31 – 43.

Binns, R., Veale, M., Van Kleek, M., & Shadbolt, N. (2017). Like Trainer, Like Bot? Inheritance of Bias in Algorithmic Content Moderation. *Social Informatics*.

Blais, M. et Martineau, S. (2007). L'analyse inductive générale: description d'une démarche visant à donner un sens à des données brutes. *Recherches qualitatives, 26*(2), 1-18. https://doi.org/10.7202/1085369ar

Bowler, L. & Knobel, C. (2014). From Cyberbullying to Well-Being: A Narrative-Based Participatory Approach to Values-Oriented Design for Social Media, *Journal of the Association for Information Science and Technology 66*(6), 1274 – 1293

Bowler, L., Knobel, C., & Mattern, E. (2015). From cyberbullying to well-being: A narrative-based participatory approach to values-oriented design for social media. *Journal of the Association for Information Science and Technology 66*(6), 1274–1293. Burch, L. (2018). "You are a parasite on the productive classes": online disablist hate speech in austere times. *Disability & Society,* 33(3), 392–415.https://doi.org/10.1080/09687599.2017.1411250

Buyse, A. (2014). Dangerous expressions: The ECHR, violence and free speech. *International & Comparative Law Quarterly, 63(*2), 491-503.

Cannard, C. (2019). Chapitre 9. Le développement social à l'adolescence: Relations aux pairs: In *Le développement de l'adolescent,* 269-299. De Boeck Supérieur. https://doi.org/10.3917/dbu.canna.2019.01.0269

Castaño-Pulgarín, S., Suárez-Betancur, N., Vega, L. & López, H. (2021). Internet, social media and online hate speech. Systematic review. *Aggression and Violent Behavior*, *58*, 101608. https://doi.org/10.1016/j.avb.2021.101608

Castets-Renard, C. (2020). Algorithmic Content Moderation on Social Media in EU Law: Illusion of Perfect Enforcement. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3535107

Chetty, N. & Alathur, S. (2018). Hate speech review in the context of online social networks. *Aggression and Violent Behavior*, *40*, 108–118. https://doi.org/10.1016/j.avb.2018.05.003

Clark, D. A., Spanierman, L. B., Reed, T. D., Soble, J. R., & Cabana, S. (2011). Documenting Weblog expressions of racial microaggressions that target American Indians. *Journal of Diversity in Higher Education, 4*(1), 39–50. https://doi.org/10.1037/a0021762

Commissie voor de evaluatie van de federale antidiscriminatiewetten, Bestrijding van discriminatie, haatboodschappen en haatmisdrijven: een gedeelde verantwoordelijkheid (2022). Available: www.unia.be/files/Evaluatiecommissie_Antidiscriminatiewetten_-_Verslag_(2022).pdf

Conseil de l'Europe. (2021, Septembre 20). Modération de contenu: Note d'orientation adoptée par le Comité directeur sur les médias et la société de l'information lors de sa 19ème séance plénière.

Cookingham, L. M., & Ryan, G. L. (2015). The Impact of Social Media on the Sexual and Social Wellness of Adolescents. *Journal of Pediatric and Adolescent Gynecology*, *28*(1), 2-5.

Cooper, K., Quayle, E., Jonsson, L., & Svedin, C. G. (2016). Adolescents and self-taken sexual images : A review of the literature. *Computers in Human Behavior, 55*, 706-716. https://doi.org/10.1016/j.chb.2015.10.003

Constantine, M. (2007). Racial Microaggressions Against African American Clients in Cross-Racial Counseling Relationships. *Journal of Counseling Psychology, 54*(1), 1–16. https://doi.org/10.1037/0022-0167.54.1.1

Cornish, D., & Clarke, R. (2017). *The reasoning criminal: Rational choice perspectives on offending* [Book]. https://doi.org/10.4324/9781315134482

Costello, M. & Hawdon, J. (2020). Hate Speech in Online Spaces. In *The Palgrave Handbook of International Cybercrime and Cyberdeviance* (p.1397–1416). *Springer International Publishing*. https://doi.org/10.1007/978-3-319-78440-3_60

Costello, M., Hawdon, J., Bernatzky, C. & Mendes, K. (2019). Social Group Identity and Perceptions of Online Hate. *Sociological Inquiry*, 89(3), 427–452. https://doi.org/10.1111/soin.12274

Costello, M., Hawdon, J., Ratliff, T., & Grantham, T. (2016). Who views online extremism? Individual attributes leading to exposure*. Computers in Human Behavior, 63*, 311–320

Costello, M., Restifo, S. J., & Hawdon, J. (2021). Viewing anti-immigrant hate online: An application of routine activity and Social Structure-Social Learning Theory. *Computers in Human Behavior*, *124*, 106927. https://doi.org/10.1016/j.chb.2021.106927

Cottee, S. (2021). Incel (E)motives: Resentment, Shame and Revenge. *Studies in Conflict & Terrorism*, *44*(2), 93114. https://doi.org/10.1080/1057610X.2020.1822589

Couturiaux, D., Young, H., Anthony, R., Page, N., Lowthian, E., Melendez-Torres, G., Hewitt, G., & Moore, G. F. (2021). Risk Behaviours Associated with Dating and Relationship Violence among 11–16 Year Olds in Wales: Results from the 2019 Student Health and Wellbeing Survey. *International Journal of Environmental Research and Public Health*, *18*(3), 1192. https://doi.org/10.3390/ijerph18031192

Cramer, R. J., Cacace, S. C., Sorby, M., Adrian, M. E., Kehn, A., & Wilsey, C. N. (2022). A Psychometric Investigation of the Hate-Motivated Behavior Checklist. *Journal of Interpersonal Violence, 38*(7-8), p. 5638-5660. https://doi.org/10.1177/08862605221127196

Crespi, I., & Hellsten, L. M. (2022). Cyberviolence and the digital experience: reflections on a problematic issue for youth*. International Review of Sociology, 32*(3), 391–399. https://doi.org/10.1080/03906701.2022.2133404

De Gregorio, G. (2020). Democratising online content moderation: A constitutional framework. *Computer Law & Security Review*, *36*, 105374. https://doi.org/https://doi.org/10.1016/j.clsr.2019.105374

Dekker, A., Wenzlaff, F., Daubmann, A., Pinnschmidt, H., & Briken, P. (2019). (Don't) Look at Me ! How the Assumed Consensual or Non-Consensual Distribution Affects Perception and Evaluation of Sexting Images. *Journal of Clinical Medicine*, *8*(5), 706. https://doi.org/10.3390/jcm8050706

Delanote, B. Peeters, & I. Van De Woesteyne (Eds.), *Digitalisering - XLVIe Postuniversitaire cyclus Willy Delva* (pp. 389-414). Wolters Kluwer.

Dias, T. (2020). Content Moderation Technologies: Applying Human Rights Standards to Protect Freedom of Expression. *Human Rights Law Review*, *20*(4), 607-640. https://doi.org/10.1093/hrlr/ngaa032

Döring, N., & Mohseni, M. R. (2019). Fail videos and related video comments on YouTube: a case of sexualization of women and gendered hate speech? *Communication Research Reports, 36*(3), 254–264. https://doi.org/10.1080/08824096.2019.1634533

Drootin, A. (2021). '"Community Guidelines": The Legal Implications of Workplace Conditions for Internet Content Moderators'. *Fordham Law Review*, 90(3).

Durán, M., & Rodríguez-Domínguez, C. (2023). Sending of Unwanted Dick Pics as a Modality of Sexual Cyber-Violence: An Exploratory Study of Its Emotional Impact and Reactions in Women. *Journal of Interpersonal Violence*, *38*(5-6), 5236-5261. https://doi.org/10.1177/08862605221120906

Estanyol, E., Montaña, M., Fernández-de-Castro, P., Aranda, D., & Mohammadi, L. (2023). Competencias digitales de la juventud en España: Un análisis de la brecha de género. *Comunicar, 31*(74), 113–123. https://doi.org/10.3916/C74-2023-09

Fino., A. (2020). Defining Hate Speech: A Seemingly Elusive Task. *Journal of International Criminal* Justice 18(1) 31–57, https://doi.org/10.1093/jicj/mqaa023

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in Text. *ACM Computing Surveys, 51*(4), 1–30.

Gagliardone, I, Danit, G., Thiago, A., Martinez, G. (2015). *Countering Online Hate Speech*, *Unesco 2015*. https://unesdoc.unesco.org/ark:/48223/pf0000233231.

Gámez-Guadix, M., Mateos-Pérez, E., Wachs, S., Wright, M., Martínez, J., & Íncera, D. (2022). Assessing image-based sexual abuse: Measurement, prevalence, and temporal stability of sextortion and nonconsensual sexting ("revenge porn") among adolescents. *Journal of Adolescence*, *94*(5), 789-799. https://doi.org/10.1002/jad.12064

Gangi, O. & Mathys, C.. Discours de haine en ligne : vers une plus grande compréhension des définitions et des expériences d'un public cible âgé de 15 à 25 ans en Belgique selon les caractéristiques individuelles. Ouvrage collectif CICY (forthcoming).

Gangi, O., Brassine, N., & Mathys, C. (2023). Entre discours de haine en ligne et cyberharcèlement chez un public belge de 15 à 25 ans: une distinction de fait et de droit, mais une distinction pertinente en criminologie?. *Criminologie, Forensique Et Sécurité*, 1(1). https://doi.org/10.26034/la.cfs.2023.3620

Gangi, O., Giacometti, M., & Gilen, A. (2022). Diffusion non consentie de contenus à caractère sexuel et diffusion d'images d'abus sexuels de mineurs: entre distinctions et chevauchements, quelles implications d'un point de vue légal, criminologique et psycho-social? *Revue de la Faculté de Droit de l'Université de Liège, 3*, 635-67.

Garcia, C. (2010). Conceptualization and Measurement of Coping During Adolescence: A Review of the Literature. *Journal of Nursing Scholarship*, *42*(2), 166-185. https://doi.org/https://doi.org/10.1111/j.1547-5069.2009.01327.x

Gassó, A. M., Klettke, B., Agustina, J. R., & Montiel, I. (2019). Sexting, Mental Health, and Victimization Among Adolescents: A Literature Review. *International Journal Environmental Research of Public Health*, *16*(13). https://doi.org/10.3390/ijerph16132364

Gerrard, Y. (2020). Social media content moderation: six opportunities for feminist intervention. *Feminist Media Studies*, *20*(5), 748-751. https://doi.org/10.1080/14680777.2020.1783807

Gibbons, F. X., & Gerrard, M. (1995). Predicting young adults' health risk behavior. *Journal of Personality and Social Psychology*, *69*(3), 505-517. https://doi.org/10.1037/0022-3514.69.3.505

Gill, K. (2021). 'Regulating Platforms' Invisible Hand: Content Moderation Policies and Processes'. *Wake Forest Journal of Business and Intellectual Property Law*, *21*(2), 171-212.

Gillespie, T. (2018). Custodians of the Internet. *Yale University Press*.

Ging, D. (2019). Alphas, Betas, and Incels: Theorizing the Masculinities of the Manosphere. *Men and Masculinities*, *22*(4), 638–657. https://doi.org/10.1177/1097184X17706401

Glowacz, F., & Goblet, M. (2020). Sexting à l'adolescence : des frontières de l'intimité du couple à l'extimité à risque. *Enfances, Familles, Générations*, *34*. https://doi.org/10.7202/1070310ar

Gongane, V. U., Munot, M. V., & Anuse, A. D. (2022). Detection and moderation of detrimental content on social media platforms: current status and future directions. *Social Network Analysis and Mining*, *12*(1), 129. https://doi.org/10.1007/s13278-022-00951-3

Guerra, C., Farkas, C., & Moncada, L. (2018). Depression, anxiety and PTSD in sexually abused adolescents: Association with self-efficacy, coping and family support. *Child Abuse & Neglect*, *76*, 310-320. https://doi.org/https://doi.org/10.1016/j.chiabu.2017.11.013

Hall, C. (2009). Sticks and stones may break bones but will the law ever protect me? Ensuring educational access through federal prohibition of peer-on-peer harassment. *Children's Legal Rights Journal, 29*(4), 42.

Harder, S. K., Jørgensen, K. E., Gårdshus, J. P., & Demant, J. (2019). *Digital sexual violence: Image-based sexual abuse among Danish youth*. In Rape in the Nordic countries (pp. 205-223). Routledge.

Harper, C., Fido, D., & Petronzi, D. (2021). Delineating non-consensual sexual image offending: Towards an empirical approach. *Aggression and Violent Behavior*, *58*, 101547. https://doi.org/10.1016/j.avb.2021.101547

Hartwig, R. & Heckenlively, K. (2021). *Behind the mask of Facebook. A wistleblower's shocking story of Big Tech bias and censorship*, New York, Skyhorse Publishing. 318.

Hawdon, J., Oksanen, A., & Räsänen, P. (2017). Exposure to Online Hate in Four Nations: A Cross-National Consideration. *Deviant Behavior*, *38*(3), 254–266. https://doi.org/10.1080/01639625.2016.1196985

Hawdon, J., Oksanen, A., & Räsänen, P. (2015). Online Extremism and Online Hate: Exposure Among Adolescents and Young Adults in Four Nations. *Nordicom Information*, *37*, 29-37.

Henry, N., Flynn, A., & Powell, A. (2019). Image-based sexual abuse: Victims and perpetrators. *Trends and Issues in Crime and Criminal Justice*, *572*, 1–19.

Henry, N., McGlynn, C., Flynn, A., Johnson, K., Powell, A., & Scott, A. (2021). *Image-based Sexual Abuse: A Study on the Causes and Consequences of Non-consensual Nude or Sexual Imagery*. Routledge.

Hirschi, T. (1969). Causes of delinquency. *Berkeley, CA: University of California Press.*

Holmes, L., Nilssen, A. R., Cann, D., & Strassberg, D. S. (2021). A sex-positive mixed methods approach to sexting experiences among college students. *Computers in Human Behavior*, *115*, 106619. https://doi.org/10.1016/j.chb.2020.106619

Jacks, W., & Adler, J. (2016). A Proposed Typology of Online Hate Crime. *Open Access Journal of Forensic Psychology*, *7*, 64-89.

Jane, E. A. (2018). Systemic misogyny exposed: Translating Rapeglish from the Manosphere with a Random Rape Threat Generator. *International Journal of Cultural Studies, 21*(6), 661–680. https://doi.org/10.1177/1367877917734042

Kaye, D. (2019). *Speech Police: The Global Struggle to Govern the Internet*. Columbia Global Reports. https://doi.org/10.2307/j.ctv1fx4h8v

Keipi, T., Näsi, M., Oksanen, A., & Räsänen, P. (2017). *Online hate and harmful content: Cross-national perspectives*. London: Routledge.

Kirchengast, T. & Crofts, T. (2018). The legal and policy contexts of 'revenge porn' criminalisation: the need for multiple approaches, Oxford University Commonwealth Law Journal 19(1), 1 – 29.

Kothari, R., Key, R., Lawrenson, J., Squire, T., Franhma, F. & Underwood, A. (2021). Understanding Risk of Suicide among perpetrators who view child sexual abuse material (CSAM). *Journal of Forensic and Legal Medicine, 81,* 102188. https://doi.org/10.1016/j.jflm.2021.102188.

Krause, N., Ballaschk, C., Schulze-Reichelt, F., Kansok-Dusche, J., Wachs, S., Schubarth, W., & Bilz, L. (2021). "Ich lass mich da nicht klein machen!" Eine qualitative Studie zur Bewältigung von Hatespeech durch Schüler/innen. *Zeitschrift für Bildungsforschung*, *11*(1), 169–185. https://doi.org/10.1007/s35834-021-00291-w

Küpper, B., Wolf, C., & Zick, A. (2010). Social Status and Anti-Immigrant Attitudes in Europe: An Examination from the Perspective of Social Dominance Theory. *International Journal of Conflict and Violence (IJCV)*, *4*(2), 205-219. https://doi.org/10.4119/IJCV-2826

Lawless, R., Robbennolt, J. & Ulen, T. (2016). *Empirical Methods in Law (2nd edition)*. Aspen Publishing.

Lee, C., Czaja, S., Moxley, J.., Sharit, J., Boot, W., Charness, N., & Rogers, W. (2019). Attitudes Toward Computers Across Adulthood From 1994 to 2013. *The Gerontologist*, *59*(1), 22-33. https://doi.org/10.1093/geront/gny081

Madigan, S., Ly, A., Rash, C. L., Van Ouytsel, J., & Temple, J. R. (2018). Prevalence of Multiple Forms of Sexting Behavior Among Youth: A Systematic Review and Meta-analysis. *JAMA Pediatrics*, *172*(4), 327-335. https://doi.org/10.1001/jamapediatrics.2017.5314

Magits, W., Van Den Elsacker, D., De Pauw, H., Michielsen, M., Aerts, Y., Bachrouri, A., & Foré, F. (2023, September 4). *Seksuele vorming verdient meer dan een vluggertje.* De Morgen. https://www.demorgen.be/meningen/seksuele-vorming-verdient-meer-dan-een-vluggertje~b807f734/

Margaret, K., Ngigi, S., & Mutisya, S. (2018). Sources of Occupational Stress and Coping Strategies among Teachers in Borstal Institutions in Kenya. *Edelweiss: Psychiatry Open Access*. https://doi.org/10.33805/2638-8073.111

McGlynn, C., Johnson, K., Rackley, E., Henry, N., Gavey, N., Flynn, A., & Powell, A. (2020). 'It's Torture for the Soul': The Harms of Image-Based Sexual Abuse. *Social & Legal Studies*, *30*(4), 541-562. https://doi.org/10.1177/0964663920947791

Margaryan, A., Littlejohn, A., & Vojt, G. (2011). Are digital natives a myth or reality? University students' use of digital technologies. *Computers & Education*, *56*(2), 429-440.

Mchangama, J., & Alkiviadou, N. (2021). Hate speech and the European Court of Human Rights: Whatever Happened to the Right to Offend, Shock or Disturb? *Human Rights Law Review*, *21*(4), 1008-1042.https://doi.org/10.1093/hrlr/ngab015

McPhail, B. A. (2002). Gender-bias h-Hate Crimes: A Review. *Trauma, Violence & Abuse*, *3*(2), 125–143. https://doi.org/10.1177/15248380020032003

Meechan-Rogers, R., Jones, C. B., & Ward, N. (2021). *Image-Based Sexual Abuse: An LGBTQ+Perspective*. In A. Powell, A. Flynn, & L. Sugiura (Eds.), The Palgrave Handbook of Gendered Violence and Technology (pp. 297-318). Springer International Publishing. https://doi.org/10.1007/978-3-030-83734-1_15

Meyer, D. (2010). Evaluating the Severity of Hate-motivated Violence: Intersectional Differences among LGBT Hate Crime Victims. *Sociology, 44*(5), 980–995. https://doi.org/10.1177/0038038510375737

Moule, R., Decker, S. & Pyrooz, D. (2017), Technology and conflict: Group processes and collective violence in the Internet era. *Crime, Law and social change 68* (1-2), 47 – 73.

Ortiz, S. M. (2019). "You can say I got desensitized to it": How men of color cope with everyday racism in online gaming. *Sociological Perspectives, 62*(4), 572–588. https://doi.org/10.1177/0731121419837588

Ortiz, S. M. (2021). Racists without racism? From colourblind to entitlement racism online. *Ethnic and Racial Studies, 44*(14), 2637–2657. https://doi.org/10.1080/01419870.2020.1825758

Paasch-Colberg, S. & Strippel, C. (2021). "The Boundaries are Blurry…": How Comment Moderators in Germany See and Respond to Hate Comments'. *Journalism Studies*, 23. 224.

Patel, U., & Roesch, R. (2022). The Prevalence of Technology-Facilitated Sexual Violence: A Meta-Analysis and Systematic Review. *Trauma, Violence, & Abuse*, *23*(2), 428-443. https://doi.org/10.1177/1524838020958057

Pedersen, W., Bakken, A., Stefansen, K., & von Soest, T. (2022). Sexual Victimization in the Digital Age: A Population-Based Study of Physical and Image-Based Sexual Abuse Among Adolescents. *Archives of Sexual Behavior, 52*(1), 399-410. https://doi.org/10.1007/s10508-021-02200-8

Petersen, J. (2021). Uitingsmisdrijven in digitale tijden: het materieel strafrecht retweeted. In M.

Powell, A., Henry, N., Flynn, A., & Scott, A. J. (2019). Image-based sexual abuse: The extent, nature, and predictors of perpetration in a community sample of Australian residents. *Computers in Human Behavior*, *92*, 393-402. https://doi.org/https://doi.org/10.1016/j.chb.2018.11.009

Powell, A., Scott, A., Flynn, A., & Henry, N. (2020). *Image-based sexual abuse: An international study of victims and perpetrators*. https://doi.org/10.13140/RG.2.2.35166.59209

Prensky, M. (2001), Digital Natives, Digital Immigrants Part 1. *On the Horizon, 9*(5), 1-6. https://doi.org/10.1108/10748120110424816

Pöyhtäri,R. (2014). 'The Limits of Hate Speech and Freedom of Speech on Moderated New Websites in Finland, Sweden, the Netherlands and the UK'. *Annales Series Historia et Sociologia, 24*. 513.

Racolța, R., & Verteș-Olteanu, A. (2019). Freedom of Expression. Some Considerations for the Digital Age. *A Journal of Social and Legal Studies*. 7-16.

Reichelmann, A., Hawdon, J., Costello, M., Ryan, J., Blaya, C., Llorent, V., Oksanen, A., Räsänen, P., & Zych, I. (2021). Hate Knows No Boundaries: Online Hate in Six Nations. *Deviant Behavior, 42*(9), 1100–1111. https://doi.org/10.1080/01639625.2020.1722337

Reinders Folmer, C. (2016). Social motives. In T*he SAGE Encyclopedia of Theory in Psychology.*

Roberts, S. (2021). *Behind the screen. Content moderation in the shadows of social media*, New Haven-London, Yale University Press, 266.

Rocque, M. (2015). The lost concept: The (re)emerging link between maturation and desistance from crime. *Criminology & Criminal Justice, 15*(3), 340–360. https://doi.org/10.1177/1748895814547710

Ryan, D. (2018). European remedial coherence in the regulation of non-consensual disclosures of sexual images. *Computer Law & Security Review 34*(5), 1053 – 1076.

Saha, K., Chandrasekharan, E., & De Choudhury, M. (2019). Prevalence and Psychological Effects of Hateful Speech in Online College Communities. *Proc Association for Computing Machinery Web Sci Conf*, *2019*, 255-264. https://doi.org/10.1145/3292522.3326032

Salminen, J., Hopf, M., Chowdhury, S. A., Soon-gyo Jung, Almerekhi, H., & Jansen, B. J. (2020). Developing an online hate classifier for multiple social media platforms. *Human-Centric Computing and Information Sciences, 10*(1), 1–34. https://doi.org/10.1186/s13673-019-0205-6

Schweppe, J., & Perry, B. (2022). A continuum of hate: delimiting the field of hate studies. *Crime, Law, and Social Change, 77*(5), 503–528. https://doi.org/10.1007/s10611-021-09978-7

Seering, J., (2020). Reconsidering Community Self-Moderation: the Role of Research in Supporting Community-Based Models for Online Content Moderation. *Proceedings of the Association of Computing Machining Human-Computer Interaction, 4* (CSCW2), 1-28.

Setty, E. (2020). *Risk and Harm in Youth Sexting Culture : Young People's Perspectives* (1st edition). Routledge. https://doi.org/10.4324/9780429277344

Siapera, E. (2022). AI Content Moderation, Racism and (de)Coloniality. *International Journal of Bullying Prevention 4*, 55–65.

Park, A. (2023). *The U.S. Surgeon General Fears Social Media Is Harming the 'Well-Being of Our Children'*. Time Magazine. Accessible: https://time.com/6282893/surgeon-general-vivek-murthy-interview-social-media/

Persily, N., & Tucker, J. (Eds.). (2020). *Social Media and Democracy: The State of the Field, Prospects for Reform* (SSRC Anxieties of Democracy). Cambridge: Cambridge University Press. doi:10.1017/9781108890960

Soral, W., Bilewicz, M., & Winiewski, M. (2018). Exposure to hate speech increases prejudice through desensitization. *Aggressive Behavior, 44*(2), 136–146. https://doi.org/10.1002/ab.21737

Sparks, B. (2022). A Snapshot of Image-Based Sexual Abuse (IBSA): Narrating a Way Forward. *Sexuality Research & Social Policy, 19*(2), 689–704. https://doi.org/10.1007/s13178-021-00585-8

Sponholz, L. (2017). Tackling hate speech with counter speech? Practices of contradiction and their effects. [Paper presented] International Conference Worlds of Contradiction, Bremen, Germany

Statbel. (2021). *Key figures 2021*. https://statbel.fgov.be/en/news/key-figures-2021

Statbel.(2023).*Diversity according to origin in Belgium*. https://statbel.fgov.be/en/themes/population/structure-population/origin#:~:text=50.3%25%20have%20the%20Belgian%20nationality,a%20foreign%20first%20registered%20nationality

Sue, D., Capodilupo, C., & Holder, A (2008). Racial Microaggressions in the Life Experience of Black Americans. *Professional Psychology, Research and Practice, 39*(3), 329–336. https://doi.org/10.1037/0735-7028.39.3.329

Suzor, N., Seignior, B. & Singleton, J. (2017). Non-consensual porn and the responsibilities of online intermediaries. *Melbourne University Law Review 40*(3), 1057-1097.

Tajfel, H. (1979). Individuals and groups in social psychology. *British Journal of Social & Clinical Psychology, 18*(2), 183–190

Tarletonn, G. (2018).*Custodians of the Internet. Platforms, content moderation, and the hiddend decisions that shape social media*, New Haven-London, Yale University Press, 288 .

Thomas, D. (2006). A General Inductive Approach for Analyzing Qualitative Evaluation Data*. The American Journal of Evaluation, 27*(2), 237 -246. https://doi.org/10.1177/1098214005283748

Tisseron, S. (2007). De l'intimité librement exposée à l'intimité menacée. *VST - Vie sociale et traitements*, *93*(1), 74. https://doi.org/10.3917/vst.093.0074

Tisseron, S. (2011). Intimité et extimité. *Communications*, *88*(1), 83. https://doi.org/10.3917/commu.088.0083

Titley, G., Keen, E. & Földi, L., Starting points for combating hate speech online (Council of Europe, 2014).

Valido, A., Espelage, D. L., Hong, J. S., Rivas-Koehl, M., & Robinson, L. E. (2020). Social-Ecological Examination of Non-Consensual Sexting Perpetration among U.S. Adolescents. *International Journal of Environmental Research and Public Health*, 17(24). https://doi.org/10.3390/ijerph17249477

Van De Heyning, C. (2022). *Internet daagt de Grondwet uit: het drukpersmisdrijf herdenken als bescherming van de vrije meningsuiting.* Jenart, C., Bernaerts, J., Peeters, Y., Popelier, P., Vanheule, D., en Verbelen, V. (eds). Liber Amicorum Jan Velaers. Die Keure. 429 - 440.

Van de Heyning, C. & Walrave, M. (2023). *Online seksueel geweld: daderschap herbekeken*. In: Mussche, C. & Stevens, L. (eds.). Onderzoek en preventie van seksuele misdrijven. Larcier Intersentia. forthcoming.

Van de Weijer, S., Leukfeldt, R., & Van der Zee, S. (2020). Reporting cybercrime victimization: determinants, motives, and previous experiences. *Policing: An International Journal, 43*(1), 17-34.

Van Gool, E., Van Ouytsel, J., Ponnet, K., & Walrave, M. (2015). To share or not to share? Adolescents' self-disclosure about peer relationships on Facebook: An application of the Prototype Willingness Model. *Computers in Human Behavior, 44,* 230-239. https://doi.org/https://doi.org/10.1016/j.chb.2014.11.036

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*, 55-59. https://doi.org/https://doi.org/10.1016/j.copsyc.2020.04.007

Van Ouytsel, J., Van Gool, E., Walrave, M., Ponnet, K., & Peeters, E. (2017). Sexting: adolescents' perceptions of the applications used for, motives for, and consequences of sexting. *Journal of Youth Studies, 20*(4), 446-470. https://doi.org/10.1080/13676261.2016.1241865

Van Ouytsel, J., Walrave, M., Ponnet, K. (2018*) 'A Nuanced Account: Why Do Individuals Engage in Sexting?'*. In: Walrave, M., Van Ouytsel, J., Ponnet, K. & Temple, J.R. (eds). Sexting: motives and risk in online sexual self-presentation. Palgrave, New York, 39 – 51.

Van Ouytsel, J., Walrave, M., De Marez, L., Vanhaelewyn, B., & Ponnet, K. (2021). Sexting, pressured sexting and image-based sexual abuse among a weighted-sample of heterosexual and LGB-youth. *Computers in Human Behavior, 117*, 106630. https://doi.org/10.1016/j.chb.2020.106630

Vanwynsberghe, H., Joris, G., Waeterloos, C., Anrijs, S., Vanden Abeele, M., Ponnet, K., De Wolf, R., Van Ouytsel, J., Van Damme, K., Vissenberg, J., D'Haenens, L., Zenner, E., Peters, E., De Pauw, S., Frissen, L., Schreuer, C.(2022*). Onderzoeksrapport Apestaartjaren : de digitale leefwereld van kinderen en jongeren*. Gent: Mediaraven. https://www.mediawijs.be/nl/onderzoek/apestaartjaren

Villanti, A. C., Johnson, A. L., Ilakkuvan, V., Jacobs, M. A., Graham, A. L., & Rath, J. M. (2017). Social Media Use and Access to Digital Technology in US Young Adults in 2016. *Journal Medical Internet Research*, *19*(6), e196. https://doi.org/10.2196/jmir.7303

Vrielink, J. (2019). *Willen we meer of minder vervolging van meningsuitingen? Rechtsantropologische vragen bij hate-speechwetgeving*. In: De grenzen van het publieke debat: verdraagzaamheid en de vrijheid van meningsuiting in de democratische samenleving, Den Haag, Boom juridisch, 89-110

Wachs, S., Bilz, L., Wettstein, A., Wright, M. F., Krause, N., Ballaschk, C., & Kansok-Dusche, J. (2022). The Online Hate Speech Cycle of Violence: Moderating Effects of Moral Disengagement and Empathy in the Victim-to-Perpetrator Relationship. *Cyberpsychology, Behavior and Social Networking, 25*(4), 223–229. https://doi.org/10.1089/cyber.2021.0159

Wachs, S., Gámez-Guadix, M., Wright, M. F., Görzig, A., & Schubarth, W. (2020). How do adolescents cope with cyberhate? Psychometric properties and socio-demographic differences of a coping with cyberhate scale. *Computers in Human Behavior*, *104*, 106167. https://doi.org/10.1016/j.chb.2019.106167

Wachs, S., Wettstein, A., Bilz, L., Krause, N., Ballaschk, C., Kansok-Dusche, J., & Wright, M. F. (2022). Playing by the Rules? An Investigation of the Relationship Between Social Norms and Adolescents' Hate Speech Perpetration in Schools. *Journal of Interpersonal Violence, 37*(21-22), 21143–21164. https://doi.org/10.1177/08862605211056032

Wachs, S., Wright, M. F., Gámez-Guadix, M., & Döring, N. (2021). How Are Consensual, Non-Consensual, and Pressured Sexting Linked to Depression and Self-Harm? The Moderating Effects of Demographic Variables. *International Journal of Environmental Research and Public Health*, *18*(5), 2597. https://doi.org/10.3390/ijerph18052597

Walker, K., & Sleath, E. (2017). A systematic review of the current knowledge regarding revenge pornography and non-consensual sharing of sexually explicit media. *Aggression and Violent Behavior*, *36*, 9-24. https://doi.org/10.1016/j.avb.2017.06.010

Walrave, M., Ponnet, K., Van Ouytsel, J., Van Gool, E., Heirman, W., & Verbeek, A. (2015). Whether or not to engage in sexting: Explaining adolescent sexting behaviour by applying the prototype willingness model. *Telematics and Informatics*, *32*(4), 796-808. https://doi.org/https://doi.org/10.1016/j.tele.2015.03.008

Waseem, Z., Davidson, T., Warmsley, D. & Weber, I. (2017). *'Understanding Abuse: A Typology of Abusive Language Detection Subtasks'*. In: Association for Computational Linguistics', Proceedings of the First Workshop on Abusive Language Online (Vancouver), 78 -84.

Wieland, J. (2007). Peer-on-peer hate crime and hate-motivated incidents involving children in California's public schools: Contemporary issues in prevalence, response and prevention. *UC Davis Journal of Juvenile Law & Policy, 11*(2), 235-269.

Willard, N. (2004). *An educator's guide to cyberbullying and cyberthreats*. Center for Safe and Responsible Use of the Internet. http://cyberbully.org

Wilson, A. & Land, M. (2021). Hate Speech on Social Media: Content Moderation in Context. *52 Connecticut Law Review, 52.*

Windler, C., M. Chair, C. Long, L. Boyle & Radovic, A. (2019). Role of Moderators on Engagement of Adolescents with Depression or Anxiety in a Social Media Intervention: Content Analysis of Web-Based Interactions?. *Journal of medical Internet research Mental, 6* (9). doi: 10.2196/13467

Wulczyn, E., Thain, N. & Dixon, L. (2017). 'Ex machina: Personal attacks seen at scale'. In: Proceedings of the 26th International Conference on World Wide Web.

Yin, R. K. (2017). *Case study research and applications: Design and methods*. Los Angeles, CA: Sage.

Young, G.K. (2022). How much is too much: the difficulties of social media content moderation. *Information & Communications Technology Law, 311*(1), 1-16.

Zhang, Z., & Luo, L. (2018). Hate speech detection: a solved problem? The challenging case of Long Tail on Twitter. *Semantic Web, 10*(5), 925–945. https://doi.org/10.3233/SW-180338

9. FIGURES & TABLES

Figures

Tables

10. ANNEXES

Annex 1	Information sheet for gatekeepers for interviews on qualitative understanding
Annex 2	Poster 1 recruitment participants for interviews on qualitative understanding
Annex 3	Poster 2 recruitment participants for interviews on qualitative understanding
Annex 4	Interview guide in Dutch
Annex 5	Interview guide in French
Annex 6	Interview guide in English
Annex 7	Legal mapping on cyberviolence
Annex 8	Dutch and French version of survey
Annex 9	Visualisation of sample respondents for the survey
Annex 10	English version of the survey
Annex 11	Chi-square tests for OHS and NCII
Annex 12	E-commerce overview on exemptions of liability
Annex 13	Due diligence obligations of OSPs
Annex 14	Programme roundtable with the industry
Annex 15	Topics discussion with industry
Annex 16	Selection of OSPs
Annex 17	Criteria for analysis OSPs self-regulatory framework
Annex 18	Questionnaire for moderators
Annex 19	Questionnaire for OSPs
Annex 20	Technical tools used by OSPs and consequences on content and user
Annex 21	Overview of numbers of complaints and questions at Unia and IEWM
Annex 22	Glossary