Aug 10th, 12:00 AM

Online content moderation and the challenge of conceptualizing cyberbullying

Zeineb Trabelsi
Université Laval, zeineb.trabelsi.1@ulaval.ca

Sehl Mellouli
Université Laval, sehl.mellouli@fsa.ulaval.ca

Richard Khoury
Université Laval, richard.khoury@ift.ulaval.ca

Online content moderation and the challenge of conceptualizing cyberbullying

Completed Research

Zeineb Trabelsi
Université Laval
zeineb.trabelsi.1@ulaval.ca

Sehl Mellouli
Université Laval
Sehl.mellouli@vre.ulaval.ca

Richard Khoury
Université Laval
richard.khoury@ift.ulaval.ca

Abstract

Cyberbullying has become a serious threat to the safety of online social communities. With the increasing growth of cyberbullying incidents, online communities rely on content moderation systems to mitigate this problem. Despite the important role of human moderators to maintain the online interactions healthy, there is a notable absence of research investigating cyberbullying within this profession. To address this gap, this study aims to comprehend moderators' understanding of cyberbullying in order to develop better and more effective intervention strategies. Using a focus group method, we explored moderators' opinions about the definitions given to cyberbullying in the scientific literature. The findings of this study highlight the importance of a message's context to detect cyberbullying rather than only considering its content. Additionally, this study indicates the role of online space factors that facilitate the occurrence of cyberbullying and the importance of clear guidelines to combat cyberbullying.

Keywords

Cyberbullying, online content moderation, online community, social media, focus group.

Introduction

The rapid growth of information and communication technologies has enabled new forms of communication and content sharing. People now have the opportunity to interact with each other in different ways. Unfortunately, these interactions are not only positive. In fact, there are people who use these interactions in harmful ways. One such harmful interaction is called cyberbullying (Bauman et al., 2013; D'Souza et al., 2018). Cyberbullying is considered a major health problem and a serious social problem. It has negative impacts on individuals' lives and it can lead to a wide spectrum of psychological problems, alcoholism and drug uses, and suicide (Bauman et al., 2013; Coelho et al., 2016; Espelage & Hong, 2017; Hinduja & Patchin, 2010; Selkie et al., 2016; Sadiq et al., 2021). Consequently, several online communities have implemented intricate and complex systems for regulating content.

In fact, these content moderation systems are a necessary tool for online platforms to prevent their digital spaces from turning into hostile environments for users (De Gregorio, 2020). Content moderation systems rely on a mix of automated programs and a group of moderators, people who set the rules, oversees their enforcement, and evaluate the worth of comments (Jhaver et al., 2019). The role of human moderators in the regulation of cyberbullying is complicated because of the complexity of the concept of cyberbullying. There is considerable debate around the definition of cyberbullying, and several different definitions have been proposed (Corcoran et al., 2015; D'Souza et al., 2018; Menesini et al., 2012; Ybarra et al., 2012; Chang, 2021). These variations and inconsistencies in defining cyberbullying may be due to a range of factors, many of which are related to people's judgments about the intent to harm, repetition, and power disparity between bullies and victims (Chang, 2021), others are related to cyber specificities of cyberbullying (Slonje et al., 2013). Additionally, without an accurate classification of cyberbullying behaviors and consistent

definitions, policy and interventions are likely to be ineffective in addressing this problem. Hence, understanding and defining cyberbullying is the first step in order to frame a comprehensive preventive policy and action plan. To address this gap, we provide a comprehensive overview of the conceptual definitions of cyberbullying by investigating the moderators' perceptions regarding the criteria used in the scientific literature (imbalance of power, intention, repetition, anonymity and lack of face-to-face contact) as well as the different categories of cyberbullying. This study aims to answer the overarching question: "what are the important factors and categories to consider in conceptualizing cyberbullying?"

Related work

Traditional criteria of Cyberbullying

In recent years, scholars have been exploring the problem of cyberbullying through numerous research efforts (Hinduja & Patchin, 2010). However, there remains a lack of consensus on how to conceptualize and operationalize this phenomenon. In fact, there is a lot of confusion on what cyberbullying consists in. For instance, several scholars have studied behaviors such as "cyber-harassment", "cyber-aggression", and "cyberstalking" under the appellation of cyberbullying (Aboujaoude et al., 2015). This illustrates that there is a need to better understand this phenomenon.

Many researchers define cyberbullying as an extension of traditional in-person bullying performed via electronic forms of contacts (Englander et al., 2017; Tahmasbi & Fuchsberger, s. d.; Whittaker & Kowalski, 2015). Traditional bullying is defined as an intentionally harmful aggressive behavior that is repetitive and involves an imbalance of power between perpetrator and target (Bauman et al., 2013; Corcoran et al., 2015; Pieschl et al., 2015; Slonje & Smith, 2008). Hence, there are three criteria to consider for cyberbullying: the repetition, the intent to harm, and the power imbalance.

Repetition

Some researchers consider bullying as a repeated action against a victim over a period of time and not just one isolated incident (Hinduja & Patchin, 2010; Menesini et al., 2012; Pieschl et al., 2015). However, the virtual nature of cyberbullying makes the determination of repetition difficult in some cases (Selkie et al., 2016; Gaffney et al., 2019). For example, a single aggressive act may be forwarded, commented or liked by many people over a short period and this may lead to repeated victimization and thus satisfy an operational definition of repetition (Corcoran et al., 2015; Pieschl et al., 2015; Selkie et al., 2016; Slonje & Smith, 2008, Peter & Petermann, 2018). This has led to an open debate of whether or not repetition is a relevant component of cyberbullying (D'Souza et al., 2018; Peter & Petermann., 2018) and allows for the notion of "repetition" in operational definitions to take a different form in the cyberspace such as sharing, liking or posting harmful content (Corcoran et al., 2015).

Intent to harm

The perpetrator must have the intention to harm the victim; the bullying act cannot occur by accident, for instance as a result of misunderstanding or a bad joke (Hinduja & Patchin, 2010). Researchers have argued that this intentionality is what makes bullying an aggressive act (Menesini et al., 2012; Pieschl et al., 2015). However, in an online context, it is difficult to capture the intent of a user (Pieschl et al., 2015). For instance, Corcoran et al. (2015) recommends defining intent to harm subjectively, as the way a reasonable person would assess the perpetrator's conduct. On the other hand, Peter & Petermann. (2018) define intent by measuring repetition, reasoning that repeated acts of bullying by a perpetrator against the victim illustrate an intention to harm him.

Power of imbalance

The imbalance of power focuses on the victim's difficulty in defending themselves against bullying events (Alim & Khalid, 2019). In traditional bullying, power imbalance can refer to many factors such as the greater physical strength, verbal competence, age, intelligence, or social status of the bully (Pieschl et al., 2015). Some researchers have proposed that this criterion differs in the case of cyberbullying and could take new

forms. Specifically, some scholars have associated the power imbalance with the anonymity that stems from using screen names and avatars online (Englander et al., 2017; Peter & Petermann., 2018). They have shown that the victims feel more powerless when faced with bullying acts from anonymous users, especially popular users, than from non-anonymous ones (Menesini et al., 2012; Pieschl et al., 2015; Tahmasbi & Fuchsberger, 2018). Other researchers have characterized power imbalance as the ability of the bully to capitalize on their technological skills (Corcoran et al., 2015).

However, several researchers have discussed the roles of repetition, intentionality and imbalance of power as criteria for both traditional bullying and cyberbullying (Alim & Khalid, 2019). The necessity of including the criterion of power imbalance in the definition of cyberbullying has been debated. In fact, some researchers have adopted definitions of cyberbullying that emphasized only repetition and intent to harm as core elements. These studies consider cyberbullying as willful and repeated harm inflicted by the use of computers, cell phones, and other electronic devices (Espelage & Hong, 2017; Hinduja & Patchin, 2010). For instance, Hood & Duffy (2018) define it as the use of computers, mobile phones, and other devices to engage in deliberate, repeated, and aggressive acts to harm others.

On the other side of the debate, Bauman et al. (2013) define cyberbullying as an intentionally harmful aggressive behavior that is repetitive and involves an imbalance of power between perpetrator and target. Uludasdemir & Kucuk (2019) describe cyberbullying as aggressive and intentional acts via electronic forms of communication by a group or an individual against those who cannot defend themselves. One of the most commonly used definitions of cyberbullying in the literature was elaborated by Smith et al., (2008 p.376): *"specific form of an aggressive, intentional act carried out by a group or individual, using electronic forms of contact, repeatedly and over time against a victim who cannot easily defend him or herself"*. This definition has been adopted by several other authors (Coelho et al., 2016; Hood & Duffy, 2018; Pieschl et al., 2015).

Cyber-Specific Criteria of Cyberbullying

Some scholars challenge the notion that cyberbullying is merely an electronic extension of traditional bullying and contest the conventional three criteria of cyberbullying (D'Souza et al., 2018; Menesini et al., 2012; Chang, 2021). They point out that defining cyberbullying by using the three offline criteria ignores the nature of cyberspace, and argue that cyberbullying requires its own, separate definition that includes cyber-specific criteria (Englander et al., 2017; Menesini et al., 2012; Pieschl et al., 2015; Peter & Petermann., 20180 such as anonymity and the lack of face-to-face contact.

Anonymity

Cyberspace offers perpetrators the possibility to be anonymous by hiding their identities, which means that individuals appear in public without being identifiable such as using fake usernames, fakes addresses (Hinduja, 2008). The anonymity available in the digital world has been shown to give a bully a persistent self-confidence to attack victims by easily enabling him to hide their identity (Slonje & Smith, 2008; Bauman et al., 2013; Tahmasbi & Fuchsberger, 2018; D'Souza et al., 2018; Peter & Petermann., 2018). By contrast, victims feel more powerless when experiencing anonymous cyberbullying (Menesini et al., 2012).

Lack of face-to-face contact

Regarding the specificity of online platforms, the perpetrator has the capacity to humiliate and to defame a victim without any face-to-face contact. This lack of direct contact between the bully and his victims, ranging from lack of direct eye contact to online invisibility, is a factor that intensifies both the negative feelings for the victim and the feeling of freedom for the bully to hurt the victim (Slonje & Smith, 2008).

Cyberbullying categories

Different researches on cyberbullying have focused on the different types of cyberbullying acts. These studies have found that cyberbullying can take many different forms (Willard, 2007; Haidar et al, 2017; Mondal et al., 2018). Based on these studies, we extracted eight main forms of cyberbullying, namely

flaming, cyber-harassment, cyberstalking, denigration, trickery, masquerading, exclusion, and doxing. Our classification scheme is a mutually exclusive and collectively exhaustive set of cyberbullying. These are defined in Table 1.

Subcategory	Definitions	Papers
Flaming	Flaming is a hostile verbal behavior, including insulting and ridiculing behavior. It consists of two or more individuals fighting with each other by sending content that involves rude, offensive, or vulgar language in order to insult the other.	Haidar et al., (2017); Mondal et al., (2018)
Cyber-harassment	Cyber-harassment consists in a perpetrator sending content that involves rude, offensive, or vulgar language in order to insult the target.	Haidar et al., (2017); Mondal et al., (2018); Willard (2007)
Cyberstalking	Cyberstalking involves intense attacks and denigration that include threats or create significant fear in the target. This includes death threats, wishes for serious disease or disability, and threats of non-consensual sexual touching.	Haidar et al., (2017); Mondal et al., (2018); Kowalski et al., (2014); Willard (2007)
Denigration	Denigration is the attempt of damaging a target's reputation or interfering with the target's friendships by insulting them to a third party. This includes posting gossip and rumors demeaning to the target.	Haidar et al., (2017); Willard (2007)
Masquerading	Masquerading occurs when the perpetrator pretends to be the target and sends offensive content, which thus appears to come from the target.	Haidar et al., (2017); Willard (2007)
Trickery	Trickery involves a situation where the perpetrator posts a material about a target that contains sensitive, private, or embarrassing information. This includes forwarding private messages or personal pictures in order to embarrass the target.	Haidar et al., (2017); Mondal et al., (2018)
Exclusion	Exclusion occurs when a perpetrator tries to force out or keep out the victim from the community. It can involve actually preventing the target from joining non-private conversations or online areas, or pressuring them to leave.	Haidar et al., (2017); Mondal et al., (2018); Willard (2007)
Doxing	Doxing involves releasing identifiable, and often private, information about an individual online; can include name, phone number, email address, home address.	Houck et al., (2014)

Table 1. Cyberbullying subcategories

Methodology

In this study, we adopt a qualitative "single focus group" approach to explore moderators' attitudes toward cyberbullying. A focus group is defined as an in-depth, open-ended group discussion that explores a specific set of issues on a predefined topic (O. Nyumba et al., 2018). This method is widely used in industry to understand sensitive behaviors such as peer aggressions, sexual misconduct, and violence (Doornwaard et al., 2017; Sigursteinsdóttir et al., 2020). Many researchers argue that a focus group is a useful approach to stimulate discussions of perceptions, opinions, and thoughts around a particular issue, as it enables people to share their experiences with the group, to clarify differences between their perspectives, and to build a consensus about sensitive behaviors (Jayasekara, 2012: Doornwaard et al., 2017: Rabiee, 2004). For these reasons, a focus group seems ideally suited for providing an in-depth picture of moderators' perceptions and comprehension of cyberbullying. Moreover, the interactive component of the focus groups enables people to ponder and listen to the experiences and opinions of others (Jayasekara, 2012).

Study Design

In our work, we employ a case study design to investigate the outlined research objectives related to the conceptualization and categorization of cyberbullying behavior in online moderation. As we already

mentioned, the definition of cyberbullying proposed by Smith et al., (2008) has found widespread acceptance in the research literature, and a large number of researchers have adopted this definition in their own work (D'Souza et al., 2018). We choose to conceptualize cyberbullying under this definition and to present it to moderators in order to understand their perspective on this phenomenon. To conduct our focus group, we elaborated an open-ended questionnaire to guide the discussion session (see Table 2). The recruitment of the focus groups participants was primarily based on their experience and expertise in the area of moderating online communities and their knowledge about cyberbullying.

Question type	Question
Key	Is cyberbullying defined in a consistent way for moderation?
Key	How are cyberbullying subcategories defined?
Key	What processes are set in place to identify and manage cyberbullying behaviors on your platform/service?
Final	Do you have any further comments to add?

Table 2. Questions guide

Participants

A selection of experts from 18 different companies specialized in online moderation were contacted via email and were asked to participate in a face-to-face focus group. In addition, the selected experts were asked to publicize the focus group project at their company and to distribute the invitation. Those who agreed to participate were demanded to specify their availabilities. Then, a follow-up email was sent to remind the participants of location, time and date and to confirm their attendance. The focus group was conducted at a hotel in London, United Kingdom. In total, 33 participants attended, of whom 14 were female and 19 male. The group comprised 5 CEOs from different companies, 12 products and community managers and 16 moderators. Years of experience ranged from 1 year to 25 years (mean 5 years ± SD 53.15). The participants are from different countries; 1 from China, 2 from Denmark, 6 from Finland, 17 from The United Kingdom, 4 from The United States, 1 from Netherlands, 1 from Ireland, and 1 from India.

Data collection

Focus group discussions require a team consisting of a skilled facilitator and assistant facilitator who are well informed about the aim of this study (O. Nyumba et al., 2018). The role of the facilitator consists of prompting participants to speak by creating a comfortable environment and by encouraging people to participate (Jayasekara, 2012; Onwuegbuzie et al., 2009; Rabiee, 2004). Given the number of participants in our group, it is also important to select observers to take notes of the emergent discussions (Onwuegbuzie et al., 2009).

Our focus group lasted approximately 75 minutes. The 33 participants were split on six tables of approximately six persons each to encourage open conversations and to allow participants to get familiar with each other. Each table was assigned an observer from the research team to take notes on the discussion. We started the focus group meeting with a presentation of the definition of cyberbullying and the different subcategories of cyberbullying identified from the literature (table 1). This was followed by a series of questions about the cyber-specific criteria of cyberbullying and about community guidelines. These questions specifically covered topics such as: how moderators manage cyberbullying incidents, and how guidelines for community interaction were established, maintained, and modified by the community managers and the CEOs. To allow participants to dive deep into those topics particularly relevant to their experience, we encouraged participants to share their thoughts and gave them a period to reflect and discuss with their colleagues at the same table. Over time, common ideas and themes began to emerge in the discussions. At that point, we felt confident that a saturation point was reached and that we had covered the key elements of cyberbullying

Data analysis

There exist four main methods to collect data: transcript-based analysis, tape-based analysis, note-based analysis, and memory-based analysis (Vogt et al., 2004; O.Nyumba et al., 2018). In this study, we opted for note-based analysis, using the notes taken by the observers during the focus group (Onwuegbuzie et al., 2009). This type of analysis is considered as a suitable approach given a time limitation (Bertrand et al., 1992), as in our case we are limited by a specific time in the symposium.

The first step of the analysis is to identify the approach to be adopted for content analysis and for coding the data. Two main approaches exist, namely inductive content analysis and deductive content analysis (Moretti et al., 2011). In the inductive approach, the categories and themes are derived from the data gathered. Meanwhile, in the deductive approach, the categories are identified during the literature review and the data is mapped to these categories (Moretti et al., 2011). In this study, we used a deductive content analysis as we have already identified our main topics; cyberbullying definition, cyberbullying categorization and cyber factors of cyberbullying. The next step is to sort data into codes. The coding process was conducted manually and the focus group notes were read thoroughly multiple times. The content analysis is based on codifying key statements by the participating experts related to each topic. Then we reduce the number of codes by collapsing them into broader themes. The analysis is self-performed, and to increase the objectivity of the findings a cross-validation by the research team was carried out. Each member of the research team read through notes taken to ensure the final themes are truly reflective to the data gathered in the focus group.

Results

The experts participating in the focus group have outlined the challenges faced when moderating online communities because of the lack of clear conceptualization of cyberbullying, which may lead moderators to perceive the phenomenon differently. Five main themes were consistently raised during the discussions: (1) the cyber-specific factors that facilitate the occurrence of cyberbullying; (2) the missing context in moderation systems, which is one factor that hinders defining and detecting cyberbullying properly; (3) the definition of cyberbullying and its sub-categories; (4) the lack of education on the topic; and (5) the importance of including different stakeholders in building online community guidelines.

Cyber-factors facilitators for cyberbullying

Participants often referred to the tendency for certain people to talk or behave online in a manner that would be inappropriate during face-to-face interactions. One participant concisely expressed the sentiment, saying users write *"like they don't have a filter there for their communications"*. Likewise, perpetrators feel less accountable for cyberbullying than they might feel for face-to-face bullying. Experts also expressed their concerns about anonymity. They pointed out when a communication is anonymous, users feel less inhibited and they can post aggressive content to hurt and to bully others. Based on a personal experience of one moderator, he noticed different behaviors in Chinese platforms as the citizen identification is used to log in; *"clients have to use citizen identification to log in so they can tell if you're a child, adult, teen, etc. Can also track user behavior."*

The importance of the context

The moderators mentioned that detecting cyberbullying out of context is very hard. "*In content moderation, context is everything*". Moderators usually do not include repetition and power imbalance in their evaluation of cyberbullying because this information is not evident to extract from short conversations. They argued that, in order to identify a repeated aggressive behavior or to determine whether a victim can defend himself, they need to consider a much longer history, which most moderation systems do not provide. Some others pointed out that repetition is an important factor to differentiate between harassment and cyberbullying and they highlighted the necessity to have a reliable technical system that contains contextual features to detect and alert moderators about cyberbullying. Some moderators questioned the usefulness of including the intent to harm as a criterion of cyberbullying, noting in particular ambiguities

in its definition, such as whether the intent to harm was present in a single event or multiple events over a period. They also pointed out that the intent could be subjective. On one side, the moderator does not know the full context, on the other side, the victim's interpretation of the intention is critical. If the act is perceived as a joke then it is not considered cyberbullying.

Defining cyberbullying and subcategories

Moderators outlined the importance of having a clear definition of cyberbullying. "Definitions are important in terms of young people understanding what is acceptable on your platform." However, a lot of debate centered around the definition of cyberbullying during the focus group. Some participants think that the definition of cyberbullying from Smith et al., (2008 p.376) is misleading. "How can we get people talking about it if the definition is misleading?"

Most notably, the scholarly definition of cyberbullying adopted in this study is characterized by a power imbalance between the victim and the perpetrator. Often the victim is described as 'weak' and unable to defend himself or herself. Some experts disagreed with this definition, saying "*the bullies are not strong and the bullied are not weak*". They suggest defining "*a bully as a person who habitually seeks to harm or intimidate those whom they perceive as vulnerable*". On the other hand, experts highlighted the high quality of the subcategories definitions of cyberbullying. There was widespread agreement about the helpfulness of this categorization in identifying what type of behaviors are bad and need to be filtered. While our classification is based on the assumption that different categories are mutually exclusive and collectively exhaustive, some moderators pointed out that the classification is not collectively exhaustive. They suggested adding other behaviors as sub-categories of cyberbullying, including stream sniping, trolling, and hacking.

The lack of education to empower community guidelines

There was a widespread agreement among moderators that both young people and adults lack knowledge and education in information and communication technologies (ICT). They mentioned that people might share the passwords of their profiles and of their emails with others. More specifically, moderators argued that young people who lack education in technology are more likely to engage in inappropriate behaviors in online communities. Consequently, moderators insist on the important role of parents and schools to monitor the children's use of ICT, and on the importance of involving parents and school administrators when defining community guidelines.

Community guidelines are a shareable responsibility

Experts stated the importance of developing shareable policies and guidelines on cyberbullying by including and getting input from different stakeholders, notably parents, young people, schools and governments. They argue that creating policies to regulate the use of social media are not the exclusive responsibility of the governments and industries. *"Government alone cannot solve this problem. I think what we found is that the industry is really keen to look at this issue and to play their part. There are actually a lot of voices that need to be listened to when it comes to the government's internet safety strategies. […] One of the most important are young people."* Without moderators and clearly defined guidelines, the discussions users engage in online have the tendency to digress. The experts agreed that there is a lack of best practices to help guide the creation of these guidelines. Encouraging the community to participate in building the guidelines could be a good way to encourage users to abide by them. *"Do you see it as the game platforms' responsibility to educate users on appropriate behavior?" "I see it as everyone's responsibility. I don't see it as the industry's sole responsibility."*

Moreover, making the community guidelines a part of the user experience and integrating them to the platform design were consistently identified as playing an important role in decreasing cyberbullying. *"It is important to make these community guidelines front and center. [….] When the user has to read it and to click to agree with that, so there is a constant reminder what the community is about".*

Discussion

In the cyberbullying literature, there often appears to be a tension between the theoretical conceptualization of cyberbullying by researchers and the practical limitations around measuring this behavior (D'Souza et al., 2018; Gaffney et al., 2019; Slonje et al., 2013). Reflecting on our findings, it appears the typical definition of cyberbullying as *"a specific form of an aggressive, intentional act carried out by a group or individual, using electronic forms of contact, repeatedly and over time against a victim who cannot easily defend him or herself" (Smith et al. 2008 p.376)* may not be as relevant in online moderation and entails a large debate.

The first definitional issue is repetition. Some participants highlighted the importance of repetition in distinguishing cyberbullying from other forms of aggression, such harassment. However, there is no full agreement on the status of repetition as a core criterion. In fact, the contextual features of cyberspace mean that such a criterion is not easily applied. Thus, moderators do not consider that repetition is a key element in defining cyberbullying. Furthermore, participants pointed out that a single incident of cyberbullying may have a significant impact on the victim. This is in line with Mehari & Doty (2021). Our findings also are consistent with D'Souza et al., (2018) who finds that repetition is not a crucial element in defining workplace cyberbullying.

The second definitional issue is power imbalance. The traditional definition of power imbalance describes the victim as "weak" (Alim & Khalid, 2019; Smith et al., 2008; Slonje et al., 2013). In the online environment, this imbalance could be described by higher technical skills in using ICT (Menesini & Nocentini, 2009; Englander et al., 2017; Peter & Petermann., 2018). However, the results of the study suggest that imbalance of power cannot be defined in terms of the inability of the victim to defend himself or in terms of higher technical skills of the perpetrator. Moderators suggested changing the definition of power imbalance.

The third definitional issue is intent to harm. As Corcoran et al. (2015) point out, and in line with the present paper, participants generally believed that intent to harm is subjective. In fact, the receiver's interpretation rather than the intent of the sender determines whether a communication constitutes cyberbullying. Many participants pointed out the prevalence of unintentional harm, which occurs when a receiver is hurt by a message or posting even though the sender did not have a malicious intent. Moreover, the result of the study highlighted that a clear indicator of intent to harm is repetition. This result is consistent with some previous research (Peter & Petermann., 2018). However, the issue of intentional and unintentional harm makes this recognition particularly challenging, specifically because the detection of the repetition is limited by the contextual features of systems moderation. Overall, there has been some difficulty in reaching a clear consensus regarding the core traditional criteria. In fact, our findings do not confirm the necessity of the three traditional criteria to define cyberbullying and they support the call for a more flexible and context-inclusive definition (D'Souza et al., 2018).

Beyond these definitional factors, this research is in line with (Slonje & Smith, 2008; Bauman et al., 2013; Tahmasbi & Fuchsberger, 2018; D'Souza et al., 2018) is that the way people act online is quite uninhibited compared with their behaviors offline. The ability to hide behind fake identities or to comment to and about strangers offers the sender a sense of empowerment. Consequently, anonymity and lack of face-to-face interaction contribute to the rise of cyberbullying on social media. Further, the findings of this research suggest that policies and guidelines may be more effective in including parents, young people, schools, industries and governments. Educating young people about the use of ICT and about different forms of bullying is particularly important to empower the community's guidelines. In regard to this complexity of the conceptualization of cyberbullying, defining criteria must be kept in mind when creating measurement tools for moderation systems and intervention and prevention strategies.

Contributions to Research and Practice

This study contributes to research and practice. Research on cyberbullying has gained significant attention in the IS literature. However, this phenomenon has suffered from both conceptual and operational problems in the past. This study provides insightful information regarding cyberbullying behaviors as perceived by online content moderators. This research demonstrates that the cyber-specific criteria have a

role in mediating cyberbullying. However, other definitional components, more specifically the notion of repetition, the notion of power imbalance, and the subjective nature of intent, need further investigation. A more standardized approach to measurement in this area is also urgently needed.

In addition, this study sheds light on the categorization and prevention of cyberbullying. Specifically, it demonstrates that cyberbullying is context-dependent and varies from case to case. Thus, there is a clear need to update content moderation tools by incorporating context features in the decision-making process. Moreover, while the process of content moderation is inherently subjective, online communities need to provide moderators with clear and consistent definitions of different forms of cyberbullying. Hence, our mutually-exclusive subcategories will help moderators to precisely recognize and identify the types of cyberbullying observed in their communities. Finally, our study suggests that online platforms and policy-makers need to work closely with other stakeholders to adjust their guidelines and to provide a healthy online atmosphere for the users.

Conclusion

The policies of different companies stipulate that harassment or cyberbullying are not allowed on their platforms. However, not all the companies have clear definitions for why a specific term was chosen, and they do not disclose the criteria for determining whether a case constitutes cyberbullying. Hence, it is a difficult challenge to build consistent moderation tools in a way that discourages cyberbullying, without building sharable guidelines and effective interventions to support community moderators in the process of decision-making. This focus group research is one of the first steps in this research area aimed at developing sharable guidelines based on the proposed categorization of cyberbullying to act on cyberbullying content. To effectively assess how to build Cyberbullying policies, companies need to share the nature of responsibility for e-safety between users, parents, schools and governments.

Acknowledgments

We would like to thank the participants of the focus groups for sharing their engagements and their valuable ideas. We also thank Two Hat Security Research Corp. for funding this research and the reviewers for their indispensable and thoughtful comments. This research was made possible by the financial support of the Canadian research organization MITACS.

REFERENCES

Slonje, R., & Smith, P. K. (2008). Cyberbullying : Another main type of bullying? Scandinavian Journal of Psychology, 49(2), 147-154.

Aboujaoude, E., Savage, M. W., Starcevic, V., & Salame, W. O. (2015). Cyberbullying : Review of an Old Problem Gone Viral. Journal of Adolescent Health, 57(1), 10-18.

Jhaver, S., Birman, I., Gilbert, E., & Bruckman, A. (2019). Human-machine collaboration for content regulation: The case of Reddit Automoderator. ACM Transactions on Computer-Human Interaction (TOCHI), 26(5), 1-35.

Alim, S., & Khalid, S. (2019). Support for cyberbullying victims and actors: A content analysis of Facebook groups fighting against cyberbullying. International Journal of Technoethics (IJT), 10(2), 35-56.

Mehari, K. R., & Doty, J. L. (2021). Bullying Conceptualization in Context: Research and Practical Implications. Human Development, 65(3), 160-165.

Haidar, B., Chamoun, M., & Serhrouchni, A. (2017). Multilingual cyberbullying detection system: Detecting cyberbullying in Arabic content. In 2017 1st Cyber Security in Networking Conference (CSNet) (pp. 1-8). IEEE.

D'Souza, N., Forsyth, D., Tappin, D., & Catley, B. (2018). Conceptualizing workplace cyberbullying: Toward a definition for research and practice in nursing. Journal of nursing management, 26(7), 842-850.

Peter, I. K., & Petermann, F. (2018). Cyberbullying: A concept analysis of defining attributes and additional influencing factors. Computers in human behavior, 86, 350-366.

Corcoran, L., Mc Guckin, C., & Prentice, G. (2015). Cyberbullying or cyber aggression? A review of existing definitions of cyber-based peer-to-peer aggression. Societies, 5, 245-255.

Chang, V. (2021). Inconsistent definitions of bullying: A need to examine people's judgments and reasoning about bullying and cyberbullying. Human Development, 65(3), 144-159.

Doornwaard, S. M., den Boer, F., Vanwesenbeeck, I., van Nijnatten, C. H., Ter Bogt, T. F., & van den Eijnden, R. J. (2017). Dutch adolescents' motives, perceptions, and reflections toward sex-related internet use: results of a web-based focus-group study. The Journal of Sex Research, 54(8), 1038-1050.

O. Nyumba, T., Wilson, K., Derrick, C. J., & Mukherjee, N. (2018). The use of focus group discussion methodology: Insights from two decades of application in conservation. Methods in Ecology and evolution, 9(1), 20-32.

Gaffney, H., Farrington, D. P., Espelage, D. L., & Ttofi, M. M. (2019). Are cyberbullying intervention and prevention programs effective? A systematic and meta-analytical review. Aggression and violent behavior, 45, 134-153.

Slonje, R., Smith, P. K., & Frisén, A. (2013). The nature of cyberbullying, and strategies for prevention. Computers in human behavior, 29(1), 26-32.

Hinduja, S., & Patchin, J. W. (2010). Bullying, Cyberbullying, and Suicide. Archives of Suicide Research, 14(3), 206-221. https://doi.org/10.1080/13811118.2010.494133.

Mondal, M., Silva, L. A., Correa, D., & Benevenuto, F. (2018). Characterizing usage of explicit hate expressions in social media. New Review of Hypermedia and Multimedia, 24(2), 110-130. https://doi.org/10.1080/13614568.2018.1489001.

Kowalski, R. M., Giumetti, G. W., Schroeder, A. N., & Lattanner, M. R. (2014). Bullying in the digital age : A critical review and meta-analysis of cyberbullying research among youth. Psychological Bulletin, 140(4), 1073-1137.

Pieschl, S., Kuhlmann, C., & Porsch, T. (2015). Beware of Publicity ! Perceived Distress of Negative Cyber Incidents and Implications for Defining Cyberbullying. Journal of School Violence, 14(1), 111-132.

Jayasekara, R. S. (2012). Focus groups in nursing research : Methodological perspectives. Nursing Outlook, 60(6), 411-416.

Selkie, E. M., Fales, J. L., & Moreno, M. A. (2016). Cyberbullying Prevalence Among US Middle and High School–Aged Adolescents : A Systematic Review and Quality Assessment. Journal of Adolescent Health, 58(2), 125-133.

De Gregorio, G. (2020). Democratising online content moderation: A constitutional framework. Computer Law & Security Review, 36, 105374.

Bauman, S., Toomey, R. B., & Walker, J. L. (2013). Associations among bullying, cyberbullying, and suicide in high school students. Journal of Adolescence, 36(2), 341-350.

Hood, M., & Duffy, A. L. (2018). Understanding the relationship between cyber-victimisation and cyber-bullying on Social Network Sites : The role of moderating factors. Personality and Individual Differences, 133, 103-108.

Menesini, E., Nocentini, A., Palladino, B. E., Frisen, A., Berne, S., Ortega-Ruiz, R., Calmaestra, J., Scheithauer, H., Schultze-Krumbholz, A., Luik, P., Naruskov, K., Blaya, C., Berthaud, J., & Smith, P. K. (2012). Cyberbullying Definition Among Adolescents : A Comparison Across Six European Countries. Cyberpsychology Behavior and Social Networking, 15(9), 455-463.

Englander, E., Donnerstein, E., Kowalski, R., Lin, C. A., & Parti, K. (2017). Defining Cyberbullying. Pediatrics, 140, S148-S151.

Uludasdemir, D., & Kucuk, S. (2019). Cyber Bullying Experiences of Adolescents and Parental Awareness : Turkish Example. Journal of Pediatric Nursing, 44, e84-e90.

Willard, N. E. (2007). Cyberbullying and cyberthreats: Responding to the challenge of online social aggression, threats, and distress. Research press.

Houck, C. D., Barker, D., Rizzo, C., Hancock, E., Norton, A., & Brown, L. K. (2014). Sexting and Sexual Behavior in At-Risk Adolescents. PEDIATRICS, 133(2), e276-e282.

Onwuegbuzie, A. J., Dickinson, W. B., Leech, N. L., & Zoran, A. G. (2009). A Qualitative Framework for Collecting and Analyzing Data in Focus Group Research. International Journal of Qualitative Methods, 8(3), 1-21.

Moretti, F., van Vliet, L., Bensing, J., Deledda, G., Mazzi, M., Rimondini, M., Zimmermann, C., & Fletcher, I. (2011). A standardized approach to qualitative content analysis of focus group discussions from different countries. Patient Education and Counseling, 82(3), 420-428.