Systematic review of determinants and consequences of bystander interventions in online hate and cyberbullying among adults

Reference:
Rudnicki Konrad, Vandebosch Heidi, Voué Pierre, Poels Karolien.- Systematic review of determinants and consequences of bystander interventions in online hate and cyberbullying among adults
Behaviour and information technology - ISSN 1362-3001 - 42:5(2023), p. 527-544
Full text (Publisher's DOI): https://doi.org/10.1080/0144929X.2022.2027013
To cite this reference: https://hdl.handle.net/10067/1855420151162165141

SYSTEMATIC REVIEW OF DETERMINANTS AND CONSEQUENCES OF BYSTANDER INTERVENTIONS IN ONLINE HATE AND CYBERBULLYING AMONG ADULTS.

Konrad Rudnicki[1] corresponding author, Heidi Vandebosch[1], Pierre Voué[2], Karolien Poels[1]

1 – Department of Communication Science, University of Antwerp, Sint-Jacobstraat 2, 2000, Antwerpen, Belgium

2 – Textgain, University of Antwerp, Lodewijk van Berckenlaan 180/2, 2140, Borgerhout, Belgium

correspondence email: kjrudnicki@gmail.com

word count: 12748

Abstract

Despite the substantial amount of literature concerning adolescent bystanders of online hate and cyberbullying, relatively little attention has been devoted to studying the same issue in adults. Similarly, the determinants of the effectiveness of different messages to support the victims or counter hate have also been understudied. The existing pieces of empirical research on these topics remained scattered and no systematic review was performed to check if there are any patterns with regard to determinants and consequences of adult bystanders intervening against hate online. To fill these gaps we performed a literature review in accordance with the guidelines of the Cochrane Collaboration Handbook for Systematic Reviews. The results of the literature search and analysis yielded three important findings. First, personal and contextual factors determining bystander action in adults largely overlap with the factors identified in adolescent populations: empathy, prior victimization, feelings of responsibility, severity, social norms, relationship with the victim and number of bystanders. Second, personal factors promoting bystander action seem to be interconnected via empathy and social norms, both of which can be facilitated through psycho-education. Third, there is a critical lack of studies on the effectiveness of different bystander interventions.

1. Introduction

The rapid advancements in communication technology are one of the defining features of the first two decades of the XXI century. The ability to instantly communicate via the Internet has given people the ability to develop their social relationships in new ways. Yet, it has also exacerbated some of the darker aspects of human nature (Brignall & Van Valley, 2005). While a hateful comment on the street can reach a handful of people at best, an angry comment on the Internet has the ability to affect millions instantly. Online hate takes many shapes and forms and grows more and more problematic (Tontodimamma et al., 2020; Waqas et al., 2019). Researchers are working tirelessly on unravelling how people attack each other online (ElSherief et al., 2018), how they react when they see hate or cyberbullying of others (Young et al., 2018) and how this phenomenon can be prevented (Blaya, 2019). They performed extensive research on this problem among children and adolescents (Allison & Bussey, 2016), where the issue is especially worrying since youth are an especially vulnerable group due to their lower ability of defending themselves from online violence and potential emotional and developmental consequences of being bullied (Blaya, 2017; Tynes et al., 2015). Unfortunately, at the same time, far less attention has been devoted to studying adult populations where the problem is dire as well.

Until the last few years most research focused on the victims and perpetrators of hate (Allison & Bussey, 2016). However, recently researchers recognized that bystanders are crucial to solving the issue of online aggression and several new studies have been published attempting to answer the questions: how do people behave when they witness hate online? This question involves a majority of the people who use the internet, as Lenhart et al. (2011) found that 88% of teenagers in the US have been witnesses of cyberbullying. The way in which bystanders choose to behave in such circumstances is vital for the outcomes on the victims and the society as a whole. Supporting the victims directly by comforting them may alleviate the deleterious effects of hate that they had been subjected to (Bastiaensens et al., 2015; Salmivalli, 2010), whereas supporting them indirectly by reporting the incidents to appropriate authorities may reduce the amount of aggressive content online and positively impact social norms of the Internet users (Anderson et al., 2014). The power of the bystander is vast, as if they join in with the perpetrators, they may encourage them to become even more aggressive and traumatize the victim even more (Bastiaensens et al., 2014; Brody & Vangelisti, 2016).

Unfortunately, despite the potential positive impact of bystander interventions most individuals choose to remain passive when witnessing online hate (Allison & Bussey, 2016). Between 50% to 90% of adolescents report that they choose to ignore when someone is being harassed online (Huang & Chou, 2010; Lenhart et al., 2011; Van Cleemput et al., 2014). Similarly, most adults also remain passive in such situations (Hayes, 2019; Henson et al., 2019). Researchers explain that intervening requires the bystander to feel some sense of connection with the victim as well as sense of safety so that they would not become victimized themselves (Obermaier et al., 2016; Rafferty & Vander Ven, 2014). This is important, because if they do not experience these feelings and remain inactive, they are more likely to become perpetrators of hate themselves (Ferreira et al., 2016). As a result, researchers in the recent years have put a lot of effort into uncovering what environments and what personality traits foster bystander intervention.

The purpose of this literature review is to present an overview on that topic and summarize the current state of research with regard to: 1) the factors that determine if an adult bystander will intervene when witnessing online aggression and 2) the factors that determine if such an intervention will be successful.

1.1. Definitions

Different forms of online aggression: online hate speech, cyberbullying, online harassment, trolling and celebrity bashing all suffer from theoretical confusion surrounding their exact definitions (Kofoed & Staksrud, 2019), which means they have to be properly defined before a literature review is performed. The broadest term used in the literature is either *online aggression* or *online harassment*. Van Royen et al. (2017, p.345) defined it as "rude, threatening or offensive content directed at others by friends or strangers and performed via electronic means." "Content" in this case can refer to anything from messages, to videos or other types of threatening behaviour like doxxing or hacking. However, the term *online harassmen*t is employed in the literature relatively rarely, and researchers focus on two more specific forms it can take: *online hate speech* (also referred to as "cyberhate" – Blaya, 2019) and *cyberbullying*. Cyberbullying is defined by most researchers broadly as a "repeated, intentional act of aggression carried out through an electronic medium against a victim who is less able to defend themselves" (Smith et al., 2008, p. 376). This definition is adapted to the online environment and follows the three core elements of traditional bullying as listed by Olweus (1998): 1) intentional harm-doing, 2) an imbalance of power and 3) repetition over time. Online hate speech is a closely related concept that differs from cyberbullying in two ways: 1) it does not necessarily require repetition over time, 2) the content of hate must address innate properties of the victim (i.e., their identity), most commonly their group belonging (Chetty & Alathur, 2018). The most common definition of hate speech lists the following personal characteristics as possible reasons of online hate: race, religion, ethnic origin, sexual orientation, disability, or gender (Johnson et al., 2019). Because of such a broad scope, the concept is fuzzy and researchers treat it as an umbrella term that is supposed to capture any *"online phenomena that involve racial hate, aggression and prejudice"* (Bliuc et al., 2018, p. 76) (for an overview of how this uncertainty is handled by society and law-makers see: Galgiardone et al., 2015). Since these concepts are relatively new and still in their theoretical infancy, Chetty & Alathur (2018, p. 109) have duly noted that: "There are no universally accepted and unique definitions of hate speech."

Because the terminological disputes about online hate speech and related concepts are still ongoing, it is necessary for researchers to adopt some but not others when they begin their studies. In this review we analyse cyberbullying and online hate together because they are the two most studied forms of online aggression. Both of them involve the key components necessary for analysing the behaviour of bystanders: 1) the intention to hurt another person, 2) power imbalance, 3) online medium, 4) visibility to a larger audience. In case of hate speech that power imbalance will always be expressed through the victims' belonging to a targeted group. In cyberbullying it can be either expressed through belonging to a targeted group or individual power imbalance of a school, family or workplace. Online hate and cyberbullying are often analysed together and efforts at designing effective interventions against online aggression may benefit from simultaneously analysing data concerning both. For example, Blaya (2019, p. 164) in her review on the topic wrote that: "Cyberhate is a form of cyberbullying."

Despite their similarities, cyberbullying and online hate are distinguishable concepts. Even though cyberbullying very often involves attacks on the victims' group identity (e.g., slurs that victimize ethnic minorities or sexual minorities), which is the defining feature of hate speech (Kowalski et al., 2020), it does not have to. It is not a defining feature of cyberbullying and there are many instances where it takes a personal form without attacks on group belonging. There are also two features of cyberbullying that are not widely recognized as central to its definition, but many researchers find them important: 1) repetition over time (Nocentini et al., 2010; DeSmet et al., 2014) and 2) peer relation between the perpetrator and the victim (Burton et al., 2013). Those features are absent for online hate. In sum, online hate and cyberbullying overlap to a relatively high extent, but still retain some key differences. Despite those differences, due to scarcity of studies on bystander interventions in online aggression and due to many researchers analysing both online hate and cyberbullying simultaneously (Blaya et al., 2019), we chose to follow in their steps and analyse them together.

These efforts with regard to bystander interventions among adolescents were recently reviewed in a structured fashion by Dominguez-Hernandez et al. (2018). The authors identified two main classes of factors that determine if a bystander is going to intervene or not: contextual factors and personal factors. Contextual factors refer to the relationships between bystanders and victims and the environment in which aggression took place. Personal factors encompass individual traits of the bystanders that predispose them to becoming upstanders. The authors found that adolescents are more likely to intervene when:

1. They are more empathic and self-efficacious (personal factors)
2. They are less morally disengaged (personal factor)
3. They have previous experiences as victims (personal factor)
4. They have some kind of relationship with the victim (contextual factor)
5. The hate incident was severe or perceived as severe (contextual factor)
6. There are fewer other witnesses (contextual factor)
7. Their social environment condemns bullying and encourages support (contextual factor)
8. The victim is actively asking for help (contextual factor)
9. They correctly evaluate the situation as an ongoing event (contextual factor)
10. They do not fear retaliation (contextual factor)

Dominguez-Hernandez et al. (2018) concluded that friendship and social environment seem to be the most important predictors of bystander intervention. When adolescents witness cyberbullying, it is crucial if they have social ties to the victim, which makes them likely to intervene. Alternatively, ties to the bully make them prone to joining in with the aggression (Bastiaensens et al., 2014, 2015; DeSmet et al., 2012, 2014, 2016; Huang & Chou, 2010; Jones et al., 2011; Machačková et al., 2013; Price et al., 2014; Thomas et al., 2012; Van Cleemput et al., 2014). In the context of fighting not only cyberbullying per se, but also online hate this is highly worrying, since in many cases racist, sexist, homophobic or otherwise exclusionary messages on the internet are addressed to anonymous strangers by other anonymous strangers. In such circumstances there are no social relationships that would push bystanders into reacting, which means that other potential encouraging factors have to be taken into careful consideration. For example, Dominguez-Hernandez et al. (2018) report that the characteristics of computer-mediated communication have the potential to inhibit or facilitate the action of bystanders. If the medium (e.g., communicator, social media feed, forum) makes it difficult to figure out if the interaction between the victim and the bully is still ongoing or already over, then bystanders are heavily discouraged from acting up (Van Cleemput et al., 2014). Furthermore, if the medium is a public domain (e.g., social media) and many other potential witnesses see the incident, the bystander effect triggers and lowers the likelihood of helping the victim, although it also lowers the likelihood of joining in with the bully (Barlinska et al., 2013).

Another extensive review of adolescent bystander response was performed by Allison & Bussey (2016). Their conclusions corroborate several findings of Dominguez-Hernandez et al. (2018). The authors point specifically to perceived severity of incidents as a predictor if bystanders will take action. Adolescents often struggle to assess whether the events they witness warrant intervention or not (Holfeld, 2014), unless those events are severe enough to leave no doubt (Obermaier et al., 2016; Patterson et al., 2015). Adults reported the same problem as well

(Shultz et al., 2014). The factor that seems to facilitate intervention, even under the conditions of uncertainty, is a direct request for help by the victim (Macháčková et al., 2013), which is an important conclusion for psycho-education, because everyone has a chance of becoming both a bystander and a victim, and knowing to ask for help may save people from trauma. It may also alleviate the effects of the bystander effect and deflection of responsibility, both of which were found to prevent adolescents from helping the victims (DeSmet et al., 2012; Huang & Chou, 2010; Macháčková et al., 2013; Van Cleemput et al., 2014). Despite a handful of particular findings, there is still insufficient research to formulate clear recommendations for policy or psycho-education with regard to the role of bystanders in online hate and cyberbullying (Dominguez-Hernandez et al., 2018). Furthermore, these results were obtained only in studies with children and adolescents and there is currently no overview of such phenomena concerning adult populations.

A review of studies about cyberbullying in adults was performed by Jenaro et al. (2018). This review focused primarily on the victims and perpetrators. With regard to bystanders it only identified a modest amount of publications that typically dealt with typologies of bystander reactions. Furthermore, it reviewed studies published until 2016, while a vast majority of studies on determinants and consequences of adult bystander interventions were published after 2017. All in all, Jenaro et al. (2018) found that in order to intervene, bystanders have to feel some kind of connection to the victim, as well as feel safe so that they do not become bullied themselves (Obermaier et al., 2016; Rafferty & Vander Ven, 2014). This echoes the conclusions by Dominguez-Hernandez et al. (2018) and Allison & Bussey (2016) in adolescents and shows that the best candidates for upstanders are the peers of the victims. However, a crucial question remains: what makes their interventions effective?

Blaya (2019) asked the same question in her review of organized intervention strategies for online hate. She found that there are almost no published rigorous evaluations that could help formulate any conclusions with regard to what makes for an effective anti-hate strategy. Instead, several types of approaches to the issue were identified:

1. Legislation efforts (examples: Brennan, 2009; Galgiardone et al., 2015)
2. Automated identification of online hate (examples: Burnap & Williams, 2016; Schmidt & Wiegand, 2017)
3. Psycho-education programs and campaigns (examples: Mwangi et al., 2016; Ranieri & Fabbro, 2016)
4. Organized counter-narrative campaigns (examples: Silverman et al., 2016; Galgiardone et al., 2015)

These methods of combating hate are strongly tied to interventions of individual bystanders, since all of them attempt one way or another at facilitating those interventions. Computer scientists and linguists design computer algorithms for detection of hate speech with the intention of flagging, removing or responding to hateful messages (Poletto et al., 2020). Counter-narrative campaigns are organized efforts of real people at intervening in as many online hate incidents as possible, in a structured fashion by using pre-designed messages (Sponholz, 2016). Finally, psycho-education campaigns attempt to teach potential bystanders what are the best ways of reacting to online hate (Gagliardone et al., 2015). However, it is difficult to scientifically support any techniques of reacting, because there are almost no studies on the issue. Blaya (2019, p.169) summed up her findings with: *"(...) up to now, although intentions are good, we have no evidence that the steps that are being undertaken are effective in preventing and reducing cyberhate."*

Currently, research on bystanders in online hate and cyberbullying is suffering from uncertainty with regard to results in adult populations and to factors that influence the effectiveness of the interventions. This is important, because research revealed that adults are also at extremely high risk of being victimized online, even up to 91% (Peluchette et al., 2015).

Between 36.2% and 68.8% of adults report witnessing online hate or cyberbullying (Alhabash et al., 2013; Selkie et al., 2016), while between 0.56% and 54.3% admit to being perpetrators (Borrajo et al., 2015; Ševčíková & Šmahel, 2009). As a result, addressing the gaps in our current knowledge about this issue is of utmost importance.

2. Methods

The present study was designed to answer two research questions via a systematic review of the literature:

1. **What factors determine if adult bystanders will intervene in situations of online aggression (in particular: hate speech and/or cyberbullying)?**
2. **What properties of those interventions determine their effectiveness?**

The systematic review was performed by following the process presented in Figure 2 and Table 1. The protocol of the systematic review was prepared in accordance with the guidelines of the Cochrane Collaboration Handbook for Systematic Reviews (Higgins et al., 2019).

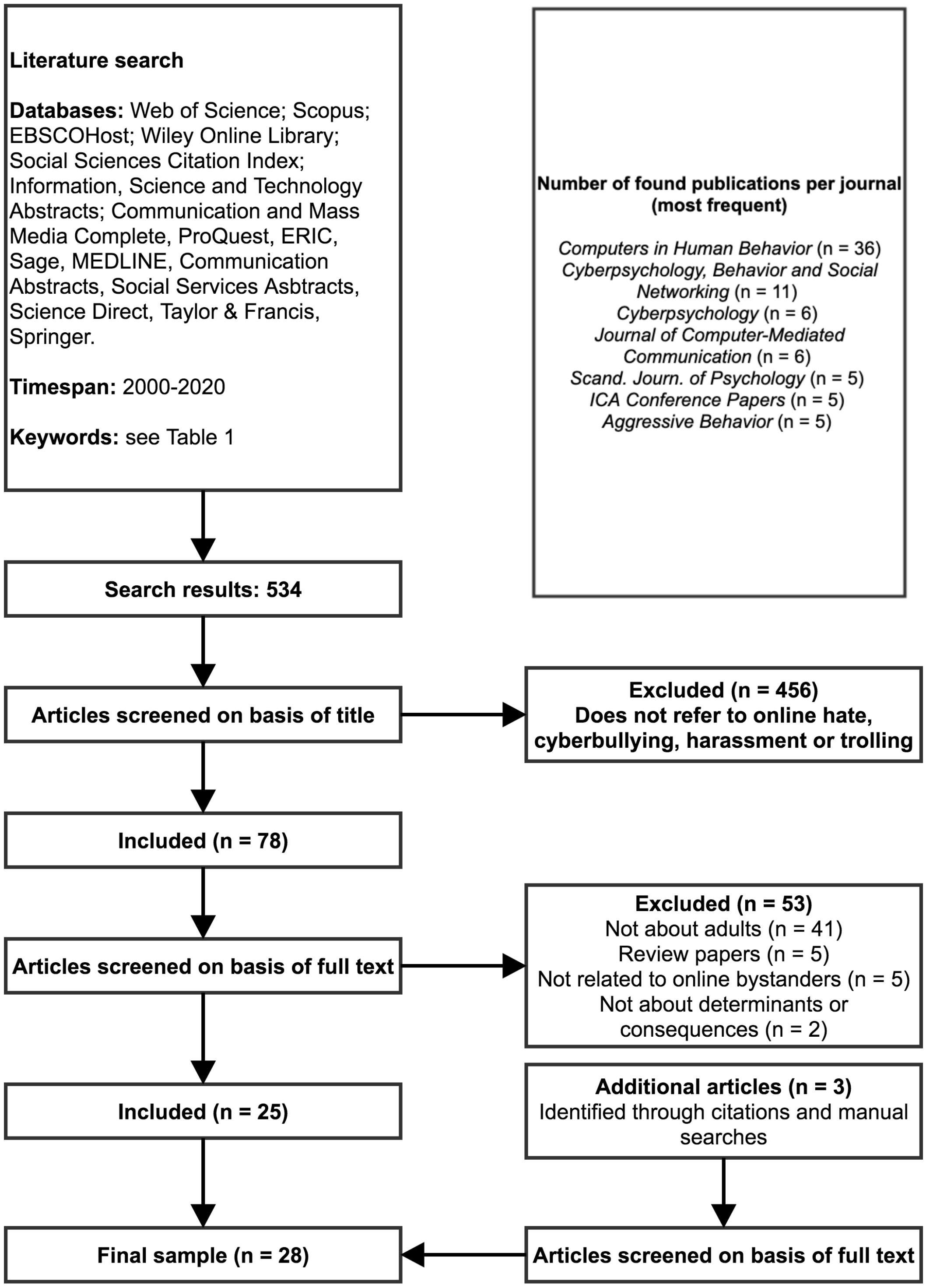
Literature search
Databases: Web of Science; Scopus; EBSCOHost; Wiley Online Library; Social Sciences Citation Index; Information, Science and Technology Abstracts; Communication and Mass Media Complete, ProQuest, ERIC, Sage, MEDLINE, Communication Abstracts, Social Services Asbtracts, Science Direct, Taylor & Francis, Springer.
Timespan: 2000-2020
Keywords: see Table 1
Number of found publications per journal (most frequent)
Computers in Human Behavior (n = 36)
Cyberpsychology, Behavior and Social Networking (n = 11)
Cyberpsychology (n = 6)
Journal of Computer-Mediated Communication (n = 6)
Scand. Journ. of Psychology (n = 5)
ICA Conference Papers (n = 5)
Aggressive Behavior (n = 5)
Search results: 534
Articles screened on basis of title
Excluded (n = 456)
Does not refer to online hate, cyberbullying, harassment or trolling
Included (n = 78)
Articles screened on basis of full text
Excluded (n = 53)
Not about adults (n = 41)
Review papers (n = 5)
Not related to online bystanders (n = 5)
Not about determinants or consequences (n = 2)
Included (n = 25)
Additional articles (n = 3)
Identified through citations and manual searches
Final sample (n = 28)
Articles screened on basis of full text

Figure 2. Flowchart of the selection process for the articles in the review.

Table 1. Keywords used in the systematic literature search.

Search nr.	Keywords	Field
1	BYSTANDER* OR WITNESS* OR THIRD-PERSON* OR UPSTANDER* OR AUDIENCE* OR PUBLIC*	**TITLE OR ABSTRACT**
	DIGITAL OR ONLINE OR SOCIAL MEDIA OR INTERNET OR CYBER* OR FACEBOOK OR TWITTER OR FORUM OR INSTAGRAM	**TITLE OR ABSTRACT**
	HATE OR HATRED	**TITLE OR ABSTRACT**
	REACT* OR RESPONSE* OR BEHAV* OR INTERVEN*	**TITLE OR ABSTRACT**
	EFFECT* OR OUTCOME* OR IMPACT* OR INFLUENCE* OR EFFICACY OR RESULT*	**TITLE OR ABSTRACT**
2	BYSTANDER* OR WITNESS* OR THIRD-PERSON* OR UPSTANDER*	**ABSTRACT**
	DIGITAL OR ONLINE OR "SOCIAL MEDIA" OR INTERNET OR CYBER* OR FACEBOOK OR TWITTER OR FORUM OR INSTAGRAM	**TITLE**
	HATE* OR HATRED* OR PREJUDICE* OR AGGRESS* OR DISCRIMINA* OR RACIS* OR MISOGYN* OR ISLAMOPHOB* OR HOMOPHOB* OR RACIAL* OR MISANDR* OR CHRISTIANOPHOB* OR TRANSPHOB* OR ETHNIC* OR RELIGIO* OR NATIONAL* OR SEXIS* OR VIOLEN* OR TROLL* OR HARASS*	**ABSTRACT**
	REACT* OR ACTI* OR RESPON* OR BEHAV* OR INTERVEN*	**ABSTRACT**
	EFFECT* OR OUTCOME* OR IMPACT* OR INFLUENCE* OR EFFICACY	**ABSTRACT**
3	BYSTANDER* OR WITNESS* OR THIRD-PERSON* OR UPSTANDER*	**ABSTRACT**

	DIGITAL OR ONLINE OR "SOCIAL MEDIA" OR INTERNET OR CYBER* OR FACEBOOK OR TWITTER OR FORUM OR INSTAGRAM	**TITLE**
	BULLY* OR CYBERBULLY*	**ABSTRACT**
	REACT* OR ACTI* OR RESPON* OR BEHAV* OR INTERVEN*	**ABSTRACT**
	EFFECT* OR OUTCOME* OR IMPACT* OR INFLUENCE* OR EFFICACY	**ABSTRACT**
4	BYSTANDER* OR WITNESS* OR THIRD-PERSON* OR UPSTANDER*	**ABSTRACT**
	DIGITAL OR ONLINE OR "SOCIAL MEDIA" OR INTERNET OR CYBER* OR FACEBOOK OR TWITTER OR FORUM OR INSTAGRAM	**TITLE**
	BULLY* OR CYBERBULLY* OR HATE OR HATRED	**ABSTRACT**
	REACT* OR ACTI* OR RESPON* OR BEHAV* OR INTERVEN*	**ABSTRACT**
	DETERMINANT* OR MODERAT* OR PREDICT*	**ABSTRACT**

The following criteria were selected for the systematic literature search:

1. The articles had to report new data (quantitative or qualitative)
2. The articles had to be focused on online aggression, such as: online hate, cyberbullying or online harassment
3. The articles had to address the determinants of bystander action or the consequences of bystander action
4. Participants of the studies had to be over the age of 18
5. The articles had to be published between 2000 and 2020

After the criteria were established, four systematic searches were performed. Databases included in the searches are reported in Figure 2. Figure 2 also presents in detail the steps taken to select the articles. Keywords used in the searches are reported in detail in Table 1. The final number of articles included in the systematic review was 29.

A majority of the selected articles used quantitative methods, although there are also 2 qualitative studies in the review. Although the search included studies that involved participants of all ages older than 18, a vast majority of them studied participants of average age in the early twenties (primarily college or university students).

The analysis of the articles was based on their full text and was performed with the following steps:

1. Identifying if the article deals with determinants or consequences of bystander behaviour
2. Identifying the core concept addressed in the study (online hate, cyberbullying, online harassment)
3. Identifying if the study addressed a specific type of hate/cyberbullying (e.g., racism, sexism, homophobia…)
4. Identifying contextual and personal factors in the articles about determinants of bystander behaviour (following the process of Dominguez-Hernandez et al., 2018)
5. Identifying different types of consequences that bystander behaviour could have as reported in the articles

3. Results

Table 2. Articles included in the systematic review. Determinants of bystander behaviour, intention of behaviour and perception of hate speech.

Authors (year)	Core concept	Type of hate	Methodology	Sample	Significant personal factors	Significant contextual factors	Outcome
Balakrishnan (2018)	Cyberbullying		Quantitative	n = 1158, M_{age} = 21, 62% female	**Fear of retaliation (-)**		Self-reported behaviour
Balakrishnan & Fernandez (2018)	Cyberbullying		Quantitative	n = 1263, M_{age} = 20.9, 62% female	**Self-esteem (+)** **Empathy (0)**		Self-reported behaviour Perception
Blackwell et al. (2018)	Online harassment		Quantitative	**Study 1** n = 160 **Study 2** n = 432		**Past negative behaviour of the victim (-)** **Intervention of others (+)**	Perception
Brody & Vangelisti (2015)	Cyberbullying		Quantitative	**Study 1** n = 265, M_{age} = 20.2, 75.1% female **Study 2** n = 379, M_{age} = 20.69, 68.6% female		**Relationship with the victim (+)** **Number of bystanders (-)** **Perceived anonymity of the bystanders (-)**	Intention

Authors (year)	Core concept	Type of hate	Methodology	Sample	Significant personal factors	Significant contextual factors	Outcome
DiFranzo et al. (2018)	Cyberbullying		Quantitative	n = 239, M_{age} = 35.14, 55.7% female	Feelings of responsibility (+) Accountability (+)	Number of bystanders (+)	Behaviour
Freis & Gurung (2013)	Cyberbullying	Homophobia	Quantitative	n = 37, 100% female	Empathy (+) Extroversion (+) Positive attitudes towards hated group (+)		Behaviour
Frischlich & Kiessler (2017)	Online hate	Racism	Quantitative	n = 391, M_{age} = 29.75, 66% female	Reactance (+)		Perception
Guo & Johnson (2020)	Online hate	Racism, Homophobia, Sexism	Quantitative	n = 368, age = 18-24, 72.6% female	Third-person effect (perceived influence of hate message on oneself) (+)		Intention
Hassan et al. (2018)	Cyberbullying		Qualitative	n = 30	Prosociality (+)	Severity (+) Number of bystanders (-) Relationship with the victim (+)	Intention
Henson et al. (2020)	Cyberbullying	Sexism	Quantitative	n = 1123, M_{age} = 20, 59% female	Self-control (+) Prior victimization (+) Social norms (+)		Self-reported behaviour

Authors (year)	Core concept	Type of hate	Methodology	Sample	Significant personal factors	Significant contextual factors	Outcome
Kazerooni et al. (2018)	Cyberbullying		Quantitative	n = 133, M_{age} = 20.8, 73% female		**Number of perpetrators (+)** **Re-shared messages (-)**	Intention
Kowalski et al. (2013)	Cyberbullying		Quantitative	n = 48, M_{age} = 18.9, 64.7% female	**Feelings of responsibility (+)** **Perceived severity (+)** **Empathic concern (+)**		Behaviour
Leonhard et al. (2018)	Online hate	Racism	Quantitative	n = 304, M_{age} = 32, 64% female		**Number of bystanders (-)** **Severity (+)**	Intention
Leung et al. (2018)	Cyberbullying		Quantitative	n = 203, Age = 12-28[2] 65% female,		**Intervention of others (+)**	Intention
Madden & Loh (2018)	Cyberbullying		Quantitative	n = 204, M_{age} = 28, 55.9% female		**Number of bystanders (-)** **Relationship with the victim (+)**	Intention
Obermaier et al. (2016)	Cyberbullying		Quantitative	**Study 1** n = 85, M_{age} = 22.35, 80.3% female **Study 2**	**Perceived severity (+)** **Feelings of responsibility (+)**	**Severity (+)** **Number of bystanders (-)**	Intention

Authors (year)	Core concept	Type of hate	Methodology	Sample	Significant personal factors	Significant contextual factors	Outcome
				n = 266, M_{age} = 23.98, 68.9% female			
Paterson et al. (2019)	Online hate	Homophobia	Quantitative	n = 465, M_{age} = 42.08, 36% female, only LGBT	**Empathy (+)** **Personal vulnerability (+)** **Prior victimization (-)**	**Victim blaming by others (-)** **Group-threat (-)**	Intention Self-reported Behaviour
Schacter et al. (2016)	Cyberbullying		Quantitative	n = 118, M_{age} = 20.55, 58% female	**Empathy (+)** **Victim blaming (-)**	**Victim behaviour (self-disclosure online) (-)**	Intention
Taylor et al. (2019)	Cyberbullying		Quantitative	**Study 1** n = 109, M_{age} = 38.15, 58% female **Study 2** n = 213, M_{age} = 37.6, 56% female	**Accountability (+)** **Empathy (+)**		Behaviour
Thacker & Griffiths (2012)	Trolling	Sexism	Qualitative	n = 125, M_{age} = 22.6,	**Prior victimization (-)**		Self-reported behaviour

				12% female			
Authors (year)	**Core concept**	**Type of hate**	**Methodology**	**Sample**	**Significant personal factors**	**Significant contextual factors**	**Outcome**
Walker et al. (2016)	Cyberbullying		Quantitative	n = 82, M_{age} = 23.96, 68.2% female	**Gender of bystander (female) (+)** **Altruism (+)**	**Gender of victim (female) (+)**	Intention
Weber et al. (2019)	Cyberbullying		Quantitative	n = 199, M_{age} = 23.85, 60% female	Sexist attitudes (+) **(-)**[3]	**Gender of victim (female) (+)**	Intention
Weber et al. (2020)	Online hate	Racism	Quantitative	n = 253, M_{age} = 43.8, 51% female	**Positive attitudes towards hated group (+)**		Behaviour
Zwillich et al. (2017)	Online hate	Racism	Quantitative	n = 132, M_{age} = 28, 70.2% female	**Feelings of responsibility (+)** **Interest in politics (+)**		Intention

Table 3. Articles included in the systematic review. Effectiveness of bystander interventions

Authors (year)	Core concept	Type of hate	Methodology	Sample	Bystander properties	Message properties	Outcome
Berman (2019)	Online hate	Racism	Quantitative	426 facebook posts	**Poster from the geopolitical West (+)**	**Personal stories (+)** **Research/policy analysis (-)** **Written posts with video (+)** **Links to websites (-)**	User engagement with counter-speech
Garland et al. (2020)	Online hate	Racism	Quantitative	1,222,240 tweets posted within 181,370 Twitter conversations		**Organized, institutional counter-speech efforts (+)** **Moderate counter-speech (+)** **Extreme counter-speech (-)**	Proportion of hate speech to counter-speech Popular support for counter-speech
High & Young (2018)	Cyberbullying		Quantitative	n = 304, M_{age} = 20.38, 82.89% female	**Experiential similarity with the victim (0)**	**Emotional comfort (+)** **Suggesting that the perpetrator could change (-)**	Victims' perceived level of support
Munger (2017)	Online hate	Racism	Quantitative	n = 242 (longitudinal)	**White, high-status male poster (congruent with the perpetrator) (+)**		Hate speech perpetration
Ozalp et al. (2020)	Online hate	Anti-semitism	Quantitative	1,232,744 tweets		**Organized, institutional counter-speech efforts (+)**	User engagement with hate speech vs. counter-speech

(+) indicates that a factor increased effectiveness of an intervention.

(-) indicates that the factor decreased effectiveness of an intervention.

3.1. Factors influencing the likelihood of bystander intervention

3.1.1. Personal factors

The analysis of the findings in the reviewed articles revealed several personal factors that appeared in more than one publication that contribute to bystanders taking action against online hate and cyberbullying.

3.1.1.1. Empathy

Perhaps the most studied personal factor determining if bystanders turn into upstanders is empathy. Paterson et al. (2019) found that empathy is a mediating factor between personal feelings of vulnerability and intentions of helping the victims. It means that participants in their study felt personally at risk of being potentially bullied in the future, which made them empathize with the victims more and, in turn, want to help them. Beyond the intention to help, Freis & Gurung (2013) found that participants higher in empathy were more likely to intervene in defence of an attacked LGBT person in a simulated Facebook environment. However, a null result with regard to empathy was reported by Balakrishnan & Fernandez (2018) who measured trait empathy in 1263 young adults. They found that empathy did not make bystanders more likely to help the victims, however in their study empathy was also not significantly altered in perpetrators either. The authors did not offer an explanation for these results and noted that a majority of their participants scored high on the empathy scale (Toronto Empathy Questionnaire – Spreng et al., 2009) putting the discriminatory power of that tool into question.

Schacter et al. (2016) reported that participants who felt less empathy for a cyberbullied victim were less likely to report an intention of helping. However, there is some terminological confusion, since the authors did not measure *empathy*, but *sympathy, compassion* or *empathic concern* – a momentary emotion of "feeling for someone" as defined by Baron-Cohen & Wheelwright (2004), which is considered to be a smaller component of empathy as a whole. The same finding as Schacter et al. (2016) and labelled as empathic concern was reported by

Kowalski et al. (2013). The distinction between trait empathy and momentary experiences of empathic concern is important, because trait empathy can be developed through lengthy psycho-education (Gehlbach, 2004; Zaki, 2014), whereas empathic concern can be manipulated via interventions. In fact, Taylor et al. (2019) whose study was analysed in our review managed to successfully demonstrate that technologically implemented empathy nudges (putting the name of the victim in the comment box: "Write to *Name"* instead of: "Write a comment," prompts asking about the feelings of people) successfully increase empathic concern for the victims and the likelihood of bystander intervention.

These results are in line with the findings in adolescent populations (Dominguez-Hernandez et al., 2018), which means that the data consistently point towards the usefulness of empathy trainings for cyberbullying and hate speech interventions programs. In fact, empathy may be trained both at its trait level for long-term effects, but also manipulated momentarily through technological nudges. As a result, it presents itself as the perfect candidate for a factor that can be used by organizational efforts in fighting hate.

3.1.1.2. Prior victimization

Several studies agree that prior victimization is an important predictor if bystanders will take action to help victims of online aggression. However, there is little agreement about the direction of that relationship. Paterson et al. (2009) found that prior LGBT hate crime victims can be more empathic towards other victims when they witness them being harassed, but only if their past experiences were indirect (they know someone who was directly victimized). If their experiences were both direct and indirect, bystanders exhibited less empathy towards the victims and more victim blaming. The authors hypothesize that such non-linear relationship may be due to the fact that extensive, compounded victimization in the past sends a message that being LGBT is not socially acceptable and that such individuals are not worthy of respect (Noelle, 2002). The more experiences, the higher chance that such beliefs become internalized and prevent the person from acting up in the future. Similarly, Thacker & Griffiths (2012) reported that people who experienced trolling in online environments were more likely to become trolls themselves.

In contrast, Henson et al. (2020) found that prior victims of sexist online harassment are consistently more likely to intervene as bystanders. In fact, the authors say that it is one of the most robust findings of their study, persistent across multiple statistical models of their data. The authors propose that victims know well when an intervention is warranted due to the memory of their own experiences. Their results corroborate the findings in adolescent populations where DeSmet et al. (2016) and Van Cleemput et al. (2014) reported that those who experienced cyberbullying themselves have a stronger intention of helping others in the same predicament.

Finally, not only the character of the relationship between prior victimization and upstanding is unclear, but also the direction of that relationship. Costello et al. (2017) examined the factors that lead people to become victims of online hate and found that actively helping victims puts a target on the bystanders' backs and makes them more likely to become attacked themselves. Therefore, we cannot say for sure if prior victims are more likely to help or do people who are likely to help in the first place just become victims more often. As a result, more research is needed to answer the question: under what circumstances does prior victimization increase the likelihood of helping victims and under what circumstances does it have an opposite effect?

3.1.1.3. Feelings of responsibility

Feeling responsible for helping is widely considered to be a mediating factor that facilitates the likelihood of helping victims of hate. DiFranzo et al. (2018) studied it as a mediator between number of witnesses and likelihood of flagging hateful posts and found it to be a significant predictor of taking action. The same approach was employed by Zwillich et al.

(2017) who were also interested in the relationship between number of bystanders and intervening on Facebook. Their results echo those of DiFranzo et al. (2018) and corroborate that feelings of responsibility are a key mediator in that relationship. Similarly, in another study feelings of responsibility were discovered to be a mediator between severity of cyberbullying and helping behaviours of bystanders (Obermaier et al., 2016). Based on these results, Taylor et al. (2019) correctly predicted that increased transparency (i.e. the presence of identifying personal information; lack of anonymity) in social media will promote action in bystanders by increasing their perceived accountability. Finally, Kowalski et al. (2013) in their experiments surveyed people about their motivations underlying the decision to help or not a bullied victim. Those participants who did not help cited their perceived lack of responsibility for the witnessed situation. Because the same results were found in several studies among adolescent populations (Dominguez-Hernandez et al., 2018) it is safe to say that perceived feelings of responsibility are a robust mediator that can be targeted as a factor that boosts likelihood of helping the victims in the future.

3.1.1.4. Perceived severity/severity/social norms

The research confirms that people react if they feel the responsibility to do so. One of the core reasons that boosts those feelings is the severity of the cyberbullying or hate incident. Severity can be operationalized as *perceived severity* (Kazerooni et al., 2018; Kowalski, 2013; Obermaier et al., 2016), which makes it a personal factor, or as *objective severity* (Hassan et al., 2018; Leonhard et al., 2018; Madden & Loh, 2018; Obermaier et al., 2016) in which case it is a contextual factor. All of the studies that account for the severity of cyberbullying and hate find that more severe cases are more likely to elicit a reaction. Because of that, it is important to answer the question: what personal factors make it more likely that a situation will be judged as severe enough to act?

In offline environments, social norms of the peers have been found to predict the likelihood of bystander intervention in multiple studies (Borsari & Carey, 2001; Brown, et al., 2010). With regard to online harassment, Henson et al. (2020) had demonstrated the same effect. In their study, participants who witnessed their peers help a sexual assault victim in the past were twice as likely to intervene in an online harassment situation. These results suggest that psycho-education with regard to bystander interventions may exhibit incremental success rates the more people participate in it, thanks to the spill-over effect of peers' social norms on those who did not participate. After all, it is the social norms we hold that affect if we judge some situations as severe enough to intervene.

3.1.1.5. Attitude towards hated group

When online aggression is based on the identity or group belonging of the victim holding implicit or explicit attitudes about that group may impact if bystanders will choose to help. Weber et al. (2020) performed an experiment in which participants were exposed to hateful, racist posts and received 5 euro that they could decide to keep or donate to a refugee aid organization. Exposition to hate speech decreased donations and that effect was mediated by implicit and explicit attitudes of the participants. In line with this idea, prejudiced attitudes towards LGBT are one of the predictors if a person is going to intervene in a homophobic cyberbullying incident (Freis & Gurung, 2018). These results are important when coupled with the studies that show how social norms of the peer group affect the helping decisions of the bystanders (Henson et al., 2020). Taken together they show that psycho-education on the norms condemning online aggression may be complimented by interventions aimed at reducing inter-group prejudice.

3.1.2. Contextual factors

3.1.2.1. Relationship with the victim

Studies on aggression in offline environments have already confirmed that people are highly more likely to defend victims with whom they have close relations (Levine & Crowther, 2008). This phenomenon extends to the online world as demonstrated by a number of studies (Brody & Vangelisti, 2016; Hassan et al., 2018; Madden & Loh, 2018). Brody & Vangelisti (2016) found this effect in a survey where participants were asked to recall a cyberbullying incident from the near past and report on their behaviour. Madden & Loh (2018) extended these results to online workplace environments and showed that if a bystander has a good working relationship with a colleague, they are more likely to defend them from bullying by a supervisor. In another study, Hassan et al. (2018) followed 30 celebrities on Instagram for 2 months and identified followers who witnessed cyberbullying (i.e., celebrity bashing) of the said celebrities. They recorded their reactions and reached out to survey them about these incidents. They found that the difference between action or inaction was often contingent on having a friendship/family relationship with the celebrity. An example reason cited by an inactive bystander read: *"I don't want to be part of it. And the person got nothing to do with me. Not in my friends and family list"* (Madden & Loh, 2018). These results closely mimic the findings in adolescent populations (Bastiaensens et al., 2014, 2015; DeSmet et al., 2012, 2014, 2016; Huang & Chou, 2010; Thomas et al., 2012) and are coherent with the fact that empathy increases the likelihood of intervention. Humans are more likely to experience empathy towards an in-group member (Forgiarini et al., 2011), which means that the closer the bystander to the victim, the more likely they will empathize with them. This could be especially true if both the victim and the bystander belong to the same victimized minority.

3.1.2.2. Number of bystanders/intervention of others

One of the best documented contextual factors that affect if bystanders will help victims of online hate is the number of other bystanders witnessing the same event (i.e., *the bystander effect*) (Brody & Vangelisti, 2015; DiFranzo et al., 2018; Hassan et al., 2018; Leonhard et al., 2018; Madden & Loh, 2018; Obermaier et al., 2016). Originally, studies on the bystander effect were sparked after the rape and murder of a woman in Queens New York. Darley & Latane (1968) were interested why none of the 38 neighbours who were witnesses, came to help the victim. They described a mechanism dubbed *diffusion of responsibility* which postulates that an increasing number of bystanders affects the likelihood of intervention by making the witnesses feel less responsible for acting up and dividing that responsibility among others (Latane & Darley, 1970). This effect is not limited to the real world setting only. A number of studies demonstrated the same phenomenon in online environments. Markey (2000) showed that posts with help requests are answered faster when they are posted on forums which have fewer active users. With regard to online hate and cyberbullying, most studies also consistently replicate the bystander effect in adults (Brody & Vangelisti, 2015; Hassan et al., 2018; Leonhard et al., 2018; Madden & Loh, 2018; Obermaier et al., 2016). In contrast to these results, one study by DiFranzo et al. (2018) reported that participants in their experiment were more likely to intervene by flagging/reporting offensive posts on social media if they were aware of an increased audience size. Furthermore, lack of information about the audience size produced similar effects as information of high audience size. The authors did not hypothesize about the possible explanations of their unexpected findings. Overwhelmingly consistent replication of the bystander effect in adults stands in contrast to relative difficulties that the researchers had in replicating it in children and adolescents, where the character of the relationship between audience size and helping seems to be non-linear (Allison & Bussey, 2016).

What may alleviate the inhibitory effect of audience size on bystander intervention are the findings of two other studies in our review. Leung et al. (2018) and Blackwell et al. (2018) showed that seeing other bystanders intervene boosts the witnesses' intention of intervening. Participants who read comments defending the victims report higher control beliefs and normative beliefs about helping them (Leung et al., 2018). Similarly, Blackwell et al. (2018)

presented their participants with online harassment exchanges where the cyberbullied person had a history of crime and manipulated if the exchanges involved someone intervening on behalf of the victim. They found that participants found the victims significantly less deserving of the bullying if someone else already defended them. These results are encouraging, because of the double-edged character that the bystander effect demonstrates. If there are more witnesses, people feel less responsible to act, but if there are more witnesses, there is a chance that someone will intervene first and facilitate the prosocial norms of other bystanders to initiate a wave of support.

3.2. Effectiveness of interventions

The reviewed literature uncovered that effectiveness of bystander interventions can be understood twofold:

1. Effectiveness with regard to the well-being of the victims
2. Effectiveness with regard to the potential at reducing similar incidents in the future (either by its impact on the perpetrator or other bystanders)

Effectiveness understood as helping the victims warrants psychological research on the properties of support messages and how they affect the emotions and cognition of the victims. In contrast, effectiveness on a societal level can be studied through the dynamics of spread of hateful messages, the engagement of Internet users with hate speech or counter speech and the incidence rates of hate and cyberbullying after the interventions in question. In this review we did not have to make decisions on whether to exclude one or the other because there is almost no published research on either. Therefore, the main result of our analysis with regard to the effectiveness of bystander interventions is a dire call to action for scientific research on the issue, since evidence-based policymaking and psycho-education cannot proceed without evidence. This is crucially important because, as High & Young (2018, p. 41) wrote:

"Understanding which messages are most effective at promoting positive outcomes is valuable for encouraging bystanders to intervene."

Some existing research informs us about the importance of counter-speech but could not have been considered for the systematic review due to methodological differences between scientific research and less conservative methodology of non-governmental organizations reporting. For example, Silverman et al. (2016) report in detail on the effectiveness of three different, organized counter-speech efforts in the US. The authors recorded over 20,000 total engagements, defined as: likes, shares, replies, retweets and comments to the counter-messaging. Unfortunately, lack of control conditions or other comparison techniques renders that data suggestive and inspiring, but inconclusive. We do not know if these campaigns ran differently would yield better or worse results. To reach scientific conclusions with regard to counter-speech, comparisons could be made between different properties of counter-messages in a controlled fashion, like in the study of Berman (2019). Alternatively, a longitudinal analysis could be performed to assess the impact of counter-speech within the windows where it was disseminated compared to windows where it was not, like in the study of Garland et al. (2020).

With regard to the effectiveness of bystander interventions on the well-being of the victims, we have identified one study (High & Young, 2018). The authors presented several types of supportive messages to self-identified victims of bullying and asked them for their assessment of the helpfulness of these messages. They found that messages containing emotional support were better at improving the affect of the victims than messages contending that the bullies had the capacity to change. Furthermore, High & Young (2018) hypothesized that bystanders who report having prior experience with being bullied are going to be more effective at supporting the victims. Instead, they found the opposite effect and observed that bystanders who did not suffer from bullying in the past caused more cognitive reappraisal in the victims. However, despite being more effective at inducing reappraisal, the same bystanders were by far the least helpful if they sent support messages that focused on the perpetrators. In other words, the victims did not like hearing anything about the perpetrators if the sender of the message did not share their pain in the past. The authors formulated directions for practice and proposed that any campaign should extensively pre-test its messages, since the identity of their senders may sometimes have counter-intuitive effect on their perception by the victims. Nevertheless, a substantial amount of additional research is needed before any directions can be formulated with satisfactory levels of certainty.

There is slightly more research concerning the effectiveness of organized bystander interventions, although we are still limited to merely four identified articles (Berman, 2019; Garland et al., 2020; Munger, 2017; Ozalp et al., 2020). Their recent publication date explains why they eluded the also very recent review by Blaya (2019). However, what lends credibility to their results and conclusions are the substantial sample sizes of online interactions that they analysed. Both Garland et al. (2020) and Ozalp et al. (2020) analysed over a million tweets in their studies.

Garland et al. (2020) focused on racist hate speech on Twitter. The authors identified two opposing groups, one called *Reconqista Germanica* which posted hateful comments about immigrants and one called *Reconquista Internet* which attempted to actively resist the first one by means of organized counter speech. Garland et al. (2020) measured the dynamics of the discourse on Twitter, mainly by quantifying the amount of engagement that bystanders exhibited with messages sent by one or the other group. Throughout their study, hateful posts and all counter speech evoked similar engagement from users, although hate seemed to attract longer exchanges of messages. However, after the counter speech group organized itself and launched coordinated efforts at countering hate, the proportion shifted, and it was counter speech that attracted more engagement. This did not reverse after the hate group organized itself in response and launched organized hate speech messaging. Overall, organized counter speech caused an initial backlash and a spike in the amount of hate, but that effect dissipated over time and the relative frequency of hate speech stabilized at lower levels than initially. Taken together, the results suggest that citizens are willing to engage with messages that support victims of hate and organized efforts at putting such messages online are demonstrably effective at increasing support for the victims of online hate.

Similarly, optimistic findings were reported by Ozalp et al. (2020) who constructed a machine learning classifier to identify anti-Semitic posts on Twitter and measured user engagement with anti-Semitic posts as well as with posts made by Jewish organizations involved in counter speech. The authors found that online hate attracted less engagement than organized counter speech. As a result, both Garland et al. (2020) and Ozalp (2020) show that being organized is a property of the bystanders that boosts their effectiveness at fighting online hate. However, a question still remains what should be the properties of the counter speech messages that these agents are posting?

Berman (2019) tried to answer that question in her doctoral dissertation. She examined if topics of facebook posts, the presence of visual aids in these posts and the geopolitical location of the poster influence the likes and shares that a counter-narrative post will gather. She discovered that including a personal story in a counter narrative is far more effective than including raw data in the form of research policies and analyses. As expected, tapping into the potential of human emotions at fostering empathy proves time and time again to be more successful than analytical arguments against aggression. Furthermore, written posts accompanied by videos gathered more engagement from the users as compared to posts that contained links to external websites. Finally, if the poster of a counter message came from the "geopolitical West," their posts attracted more attention. This result might be influenced by the population from which the audience of the messages came, since the study was performed in the US. It is likely that effectiveness of a support message will not depend on just being from "the West" but on the congruence between the perceived group identity of the authors of support messages and their recipients. This hypothesis is supported by the results of Munger (2017) who created bots on Twitter and varied the bots' social status (number of followers) and ethnicity (white vs. black). The author observed that being reprimanded by a high-status, white account on Twitter reduced the amount of racist hate speech performed by the scorned accounts for two months. This result was true for online hate spread by posters whose ethnicity was congruent with that of the bot. More research is needed on the matter, but it appears that posts made by supposed individuals whose traits align with the traits of the haters may be more effective than campaigns and messages labelled with names of institutions.

4. Discussion

This review revealed seven main factors tied to the likelihood of adult bystanders' action when witnessing online hate and cyberbullying. With regard to personal factors the literature has put the most emphasis on empathy, prior victimization, feelings of responsibility, social norms and attitudes towards vulnerable groups. On the contextual level it highlighted the importance of personal relationships between bystanders and the victims as well as the number of bystanders witnessing the incidents of online hate.

Taken together, the results with regard to the personal factors influencing bystander action in adults suggest that psycho-education and training programs may have excellent effects if they address the issue in a complex way. Trait empathy and expressions of empathic concern seem to be the most reliable predictor of bystanders taking action against online hate (Freis & Gurung, 2013; Paterson et al., 2019; Schacter et al., 2016; Taylor et al., 2019). This finding is not new, since numerous anti-hate training programs in adolescents and adults have already been designed with facilitating empathy in mind (Van Noorden et al., 2015). Reviewing the effectiveness of those programs in the context of online bystanders is difficult, since most often they focus on preventing perpetration or facilitating bystander action in real life. However, it is safe to say that any organized effort at promoting standing up against hate must incorporate empathy for others as its core concept. This conclusion follows not only from the fact that most scientific articles on the issue have been published about empathy, but also from the fact that most of the other personal factors that influence bystander action are closely tied to empathy themselves.

Our analysis identified papers that demonstrated that people are more likely to stand in defence of the victims online if their social norms dictate them to (Borsari & Carey, 2001; Brown, et al., 2010), if they feel responsible for doing so (DiFranzo et al., 2018; Kowalski et al., 2013; Obermaier et al., 2016; Zwillich et al., 2017) and if they have a positive attitude towards the victimized group (Weber et al., 2020; Freis & Gurung, 2018). All of these concepts are related to empathy. Humans exhibit strong conformity in their prosocial behaviours and expressions of empathy, which means that social norms regulate them. For example, Nook et al. (2016) demonstrated in a series of experiments that empathy can be contagious and social norms of others are the carriers of that contagion. In their study, participants expressed significantly more empathy and prosocial behaviours after they observed others doing so. Empathy is also an affective cornerstone for the attitudes towards people who are out-group members. Most of the time, humans are much more likely to feel empathy towards their own in-group members, an effect strong enough that even empathy for physical pain is going to be significantly higher for the members of their own race (Forgiarini et al., 2011). However, as empathy increases, our attitudes towards the out-groups become more inclusive. For example, Nesdale et al. (2005) showed in children that increasing empathy is correlated with more liking for people of different ethnicity. However, in addition to the relation between empathy and attitudes towards out-groups, the authors also addressed the role of social norms. They found that increasing empathy facilitates inclusive attitudes but only if the social norms within the group are inclusive. If the social norms are exclusionary, empathy could not facilitate acceptance towards other ethnicities. It follows, that empathy trainings may not be enough to root out online hate. Building an inclusive society requires scientists, policymakers, activists and psycho-educators to address social norms, especially those with regard to the acceptance of people who may be perceived as out-group. Most of the time, that would mean the vulnerable minorities. These social norms advocated by groups and attitudes presented by individuals might synergize with empathy and encourage more bystanders to take action when someone is being attacked online.

Interestingly, the discussion of contextual factors that facilitate bystander interventions may also be started by highlighting the importance of empathy. Technological interventions (nudges) were shown by Taylor et al. (2019) to have a positive effect on the empathic concern of social media users and on the likelihood that they would intervene on behalf of the victims. In media studies, the properties of the medium through which people consume content or interact with each other are considered extremely important for moderating the effects that the medium has on its user (Bandura 2009). In simple words, if someone wants to encourage a certain type of behaviour on their website, it matters what that website looks like. In line with this idea, Van Royen et al. (2017) demonstrated that "reflective interfaces" significantly decrease intention at engaging in cyberbullying. In their experiment, participants were prompted with messages before posting an aggressive comment, for example: *"This comment may be hurtful for the receiver. Are you sure to post it?"* Such interventions may seem trivial but were shown to have positive impact on behaviour. Given that empathy is already otherwise known to be a powerful force pushing people towards anti-hate behaviours and that Taylor et al. (2019) paved the way for technological empathy nudges embedded in the properties of media, we believe that this type of intervention may prove to be extraordinarily useful in the future. However, more research with regard to this particular problem has to be conducted first.

One of the most important conclusions of this literature review is a striking lack of scientific works on the effectiveness of bystander interventions. This is especially worrying, since the whole international community, including politicians, has already called for action in that regard. The Security Council of the United Nations has issued a resolution (UN Security Council Resolution 2357-2017) in which it highlights the necessity of countering online hate speech that is being used by radical and terrorist organizations. It also postulates that ICT providers, governments and non-government organizations in every country should begin coordinated efforts at countering online hate speech. Most importantly, the Security Council has also urged the members of the United Nations to monitor and evaluate the strategies undertaken in order to limit the spread of online hate speech. Unfortunately, the reports of such monitoring and evaluation are currently extremely scarce. It cannot be understated that there can be no evidence-based policies without evidence.

The few articles that were identified in this review and concerned the effectiveness of bystander interventions provide some preliminary data. It is encouraging that the results published by Garland et al. (2020) and Ozalp et al. (2020) demonstrate that organized counter-speech efforts make a visible difference in the dynamics of online discourse. However, drawing any conclusions beyond that is currently extremely difficult. Garland et al. (2020) provided some evidence for the commonly occurring recommendation for people reacting to online hate or trolls to be assertive but not aggressive. The authors write that moderate counter-speech was effective at reducing engagement with online hate, whereas extreme counter-speech produced an opposite effect. Similarly, Munger (2017), in the experiment where a bot was successfully reprimanding Twitter users for racist hate speech, used moderate messages that highlighted the potential harm to the victims. Many materials prepared by non-government organizations or other institutions mention that intervening bystanders should remain moderate in their speech (Article 19, 2018; Center for the Prevention of Hate Violence, 2020). Now there is finally some evidence for it, but more research is needed to turn this into an established claim. Additionally, Berman (2019) provided some initial data showing that personal stories are more effective than research/policy analyses when addressing online hate, and written posts with an attached video are more effective than links. However, a world of other properties that messages have remains unexplored. Are messages that contain humour more effective? Should bystanders respond to posts that are hateful enough to warrant a report or are criminal? Should the messages be posted publicly or sent directly to the victims? Are different types of responses more effective for different types of hate? These are only a few unanswered questions that must be addressed before we are confident in creating policies, guidelines for social media and normative statements for training programs.

The most important limitation of this literature review is lack of empirical data that would allow for analysing online hate and cyberbullying separately. In order to identify the most likely factors that affect bystander interventions we have taken these two phenomena together. It was possible thanks to the fact that they are conceptually very close and researchers in the field often analyse them together (Blaya 2019). However, special attention has to be devoted to any potential differences between online hate and cyberbullying which may become apparent in the future.

The intention behind countering online hate and cyberbullying is to reinforce the social norm of civil language in the cyberspace as well as promote dialogue and cooperation between people of different ethnicities, religions, etc. (Iganski, 2020). However, that also means that it is important to try preventing online hate from occurring in the first place. It is easier to prevent prejudice from sprouting before it happens rather than deconstructing norms and beliefs that people already have (Cichocka, 2020; Gärdenfors, 2003). The best type of bystander intervention is one that never had to happen, because the people who would want to harass others stop themselves in the fear of being subjected to social ostracism. That means that researchers have a moral obligation to determine what types of bystander interventions are the most effective and most ethical. Otherwise, victims of online hate and cyberbullying may be left without any guidelines what are adaptive ways of defending themselves and instead develop behaviour that would further the spread of online aggression. For example, vulnerable minorities have recently developed a phenomenon called “cancel culture,” where celebrities who express controversial opinions become aggressively ostracized and shunned (Ng, 2020). In principle, “cancel culture” may be seen as an expression of a social group delineating its boundaries by ostracizing those who publicly break them. However, the ostracism employed by the vulnerable groups “cancelling” others may sometimes border on online hate and cyberbullying itself and raises concerns about freedom of expression and public discourse (Norris, 2020). Unfortunately, without more evidence about what constitutes an effective and ethical way of countering hate we are unable to address issues like this without seemingly taking the side of the perpetrators of online hate.

Bibliography

Alhabash, S., McAlister, A. R., Hagerstrom, A., Quilliam, E. T., Rifon, N. J., & Richards, J. I. (2013). Between likes and shares: Effects of emotional appeal and virality on the persuasiveness of anticyberbullying messages on Facebook. *Cyberpsychology, Behavior, and Social Networking*, *16*(3), 175-182.

Allison, K. R., & Bussey, K. (2016). Cyber-bystanding in context: A review of the literature on witnesses' responses to cyberbullying. *Children and Youth Services Review*, *65*, 183-194.

Anderson, J., Bresnahan, M., & Musatics, C. (2014). Combating weight-based cyberbullying on Facebook with the dissenter effect. *Cyberpsychology, Behavior, and Social Networking*, *17*(5), 281-286.

Balakrishnan, V. (2018). Actions, emotional reactions and cyberbullying–From the lens of bullies, victims, bully-victims and bystanders among Malaysian young adults. *Telematics and Informatics*, *35*(5), 1190-1200.

Balakrishnan, V., & Fernandez, T. (2018). Self-esteem, empathy and their impacts on cyberbullying among young adults. *Telematics and Informatics*, *35*(7), 2028-2037.

Bandura, A. (2009). Social cognitive theory of mass communications. In: Bryant J.; Oliver MB (*eds*.). *Media effects: advances in theory and research.*

Barlinska, J., Szuster, A., & Wisniewski, M. (2015). The Role of short-and long-term cognitive empathy activation in preventing cyberbystander reinforcing cyberbullying behavior. *Cyberpsychology, Behavior, and Social Networking, 18*, 241-244.

Baron-Cohen, S., & Wheelwright, S. (2004). The empathy quotient: an investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences. *Journal of autism and developmental disorders*, *34*(2), 163-175.

Bastiaensens, S., Vandebosch, H., Poels, K., Van Cleemput, K., DeSmet, A., & De Bourdeaudhuij, I. (2015). 'Can I afford to help?' How affordances of communication modalities guide bystanders' helping intentions towards harassment on social network sites. *Behaviour & Information Technology*, *34*(4), 425-435.

Bastiaensens, S., Vandebosch, H., Poels, K., Van Cleemput, K., Desmet, A., & De Bourdeaudhuij, I. (2014). Cyberbullying on social network sites. An experimental study into bystanders' behavioural intentions to help the victim or reinforce the bully. *Computers in Human Behavior*, *31*, 259-271.

Berman, E. (2019). Evaluating the Effectiveness of Counter-Narrative Tactics in Preventing Radicalization. (Doctoral dissertation) Walden University, Minneapolis, MN, US.

Blackwell, L., Chen, T., Schoenebeck, S., & Lampe, C. (2018). When online harassment is perceived as justified. In *Twelfth International AAAI Conference on Web and Social Media*.

Blaya, C. (2017) Online racist, xenophobic and religious grounded hate speech and the experiences of the young people. *Communication presented at Children and the Youth on the Net Conference, Luxembourg* (2017, April)

Blaya, C. (2019). Cyberhate: A review and content analysis of intervention strategies. *Aggression and violent behavior*, *45*, 163-172.

Bliuc, A. M., Faulkner, N., Jakubowicz, A., & McGarty, C. (2018). Online networks of racial hate: A systematic review of 10 years of research on cyber-racism. *Computers in Human Behavior, 87*, 75-86.

Borrajo, E., Gámez-Guadix, M., Pereda, N., & Calvete, E. (2015). The development and validation of the cyber dating abuse questionnaire among young couples. *Computers in Human Behavior*, *48*, 358-365.

Brennan, F. (2009). Legislating against Internet race hate. *Information & Communications Technology Law*, *18*(2), 123-153.

Brignall III, T. W., & Van Valey, T. (2005). The impact of internet communications on social interaction. *Sociological Spectrum*, *25*(3), 335-348.

Brody, N., & Vangelisti, A. L. (2016). Bystander intervention in cyberbullying. *Communication Monographs*, *83*(1), 94-119.

Burnap, P., & Williams, M. L. (2016). Us and them: Identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science, 5*(1), 11.

Burton, K. A., Florell, D., & Wygant, D. B. (2013). The role of peer attachment and normative beliefs about aggression on traditional bullying and cyberbullying. *Psychology in the Schools, 50*(2), 103-115.

Chetty, N., & Alathur, S. (2018). Hate speech review in the context of online social networks. *Aggression and violent behavior*, *40*, 108-118.

Costello, M., Hawdon, J., & Ratliff, T. N. (2017). Confronting online extremism: The effect of self-help, collective efficacy, and guardianship on being a target for hate speech. *Social Science Computer Review*, *35*(5), 587-605.

DeSmet, A., Bastiaensens, S., Van Cleemput, K., Poels, K., Vandebosch, H., Cardon, G., & De Bourdeaudhuij, I. (2016). Deciding whether to look after them, to like it, or leave it: A multidimensional analysis of predictors of positive and negative bystander behavior in cyberbullying among adolescents. *Computers in Human Behavior*, *57*, 398-415.

DeSmet, A., Bastiaensens, S., Van Cleemput, K., Poels, K., Vandebosch, H., & De Bourdeaudhuij, I. (2012). Mobilizing bystanders of cyberbullying: an exploratory study into behavioural determinants of defending the victim. In B. K. Wiederhold & G. Riva (Eds.), *Annual review of cybertherapy and telemedicine 2012: Advanced technologies in the behavioral, social and neurosciences* (Vol. 181, pp. 58–63).

DeSmet, A., Veldeman, C., Poels, K., Bastiaensens, S., Van Cleemput, K., Vandebosch, H., & De Bourdeaudhuij, I. (2014). Determinants of self-reported bystander behavior in cyberbullying incidents amongst adolescents. *Cyberpsychology, Behavior, and Social Networking*, *17*(4), 207-215.

DiFranzo, D., Taylor, S. H., Kazerooni, F., Wherry, O. D., & Bazarova, N. N. (2018, April). Upstanding by design: Bystander intervention in cyberbullying. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1-12).

Domínguez-Hernández, F., Bonell, L., & Martínez-González, A. (2018). A systematic literature review of factors that moderate bystanders' actions in cyberbullying. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, *12*(4).

ElSherief, M., Kulkarni, V., Nguyen, D., Wang, W. Y., & Belding, E. (2018). Hate lingo: A target-based linguistic analysis of hate speech in social media. *arXiv preprint arXiv:1804.04257*.

Facebook community standards (2020, November 26) Hate speech. Retrieved from: https://www.facebook.com/communitystandards/hate_speech

Ferreira, P. C., Simão, A. V., Ferreira, A., Souza, S., & Francisco, S. (2016). Student bystander behavior and cultural issues in cyberbullying: When actions speak louder than words. *Computers in Human Behavior*, *60*, 301-311.

Forgiarini, M., Gallucci, M., & Maravita, A. (2011). Racism and the empathy for pain on our skin. *Frontiers in psychology*, *2*, 108.

Freis, S. D., & Gurung, R. A. (2013). A Facebook analysis of helping behavior in online bullying. *Psychology of popular media culture*, *2*(1), 11.

Frischlich & Kiessler (2017) I will not Hate: Reactance Moderates the Effects of Hate Speech on Prejudice. In *Conference Papers - International Communication Association 2017*, 1–15.

Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). Countering online hate speech. *UNESCO series on Internet freedom*. Paris: UNESCO.

Garland, J., Ghazi-Zahedi, K., Young, J. G., Hébert-Dufresne, L., & Galesic, M. (2020). Impact and dynamics of hate and counter speech online. (arXiv preprint) arXiv:2009.08392.

George Mwangi, C. A., Bettencourt, G. M., & Malaney, V. K. (2018). Collegians creating (counter) space online: A critical discourse analysis of the I, Too, Am social media movement. *Journal of Diversity in Higher Education*, *11*(2), 146.

Guo, L., & Johnson, B. G. (2020). Third-Person Effect and Hate Speech Censorship on Facebook. *Social Media+ Society*, *6*(2).

Hassan, S., Yacob, M. I., Nguyen, T., & Zambri, S. (2018, July). Should I Intervene? The Case of Cyberbullying on Celebrities from the Perspective of the Bystanders. In *KMICe 2018,* 382-387.

Hayes, B. E. (2019). Bystander intervention to abusive behavior on social networking websites. *Violence against women*, *25*(4), 463-484.

Henson, B., Fisher, B. S., & Reyns, B. W. (2020). There Is Virtually No Excuse: The Frequency and Predictors of College Students' Bystander Intervention Behaviors Directed at Online Victimization. *Violence Against Women*, *26*(5), 505-527.

Higgins, J. P., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2019). Cochrane handbook for systematic reviews of interventions. John Wiley & Sons.

High, A. C., & Young, R. (2018). Supportive communication from bystanders of cyberbullying: Indirect effects and interactions between source and message characteristics. *Journal of Applied Communication Research*, *46*(1), 28-51.

Holfeld, B. (2014). Perceptions and attributions of bystanders to cyberbullying. *Computers in Human Behavior, 38*, 1–7.

Huang, Y. Y., & Chou, C. (2010). An analysis of multiple factors of cyberbullying among junior high school students in Taiwan. *Computers in Human Behavior*, *26*(6), 1581-1590.

Jenaro, C., Flores, N., & Frías, C. P. (2018). Systematic review of empirical studies on cyberbullying in adults: What we know and what we should investigate. *Aggression and violent behavior*, *38*, 113-122.

Johnson, N. F., Leahy, R., Restrepo, N. J., Velasquez, N., Zheng, M., Manrique, P., ... & Wuchty, S. (2019). Hidden resilience and adaptive dynamics of the global online hate ecology. *Nature*, *573*(7773), 261-265.

Jones, S. E., Manstead, A. S., & Livingstone, A. G. (2011). Ganging up or sticking together? Group processes and children's responses to text-message bullying. *British Journal of Psychology*, *102*(1), 71-96.

Kazerooni, F., Taylor, S. H., Bazarova, N. N., & Whitlock, J. (2018). Cyberbullying bystander intervention: The number of offenders and retweeting predict likelihood of helping a cyberbullying victim. *Journal of Computer-Mediated Communication*, *23*(3), 146-162.

Kofoed, J., & Staksrud, E. (2019). 'We always torment different people, so by definition, we are no bullies': The problem of definitions in cyberbullying research. *New Media & Society*, *21*(4), 1006-1020.

Kowalski, R. M., Dillon, E., Macbeth, J., Franchi, M., & Bush, M. (2020). Racial differences in cyberbullying from the perspective of victims and perpetrators. *American journal of orthopsychiatry, 90*(5), 644.

Kowalski, R.M., Schroeder, A.N., Smith, C.A. (2013). Bystanders and their willingness to intervene in cyber bullying situations. In R. Hanewald (Ed.), *From cyber bullying to cyber safety: Issues and approaches in educational contexts* (pp. 77-100), Nova Science Publishers, New York.

Lenhart, A., Madden, M., Smith, A., Purcell, K., Zickuhr, K., & Rainie, L. (2011). Teens, Kindness and Cruelty on Social Network Sites: How American Teens Navigate the New World of" Digital Citizenship". *Pew Internet & American Life Project*.

Leonhard, L., Rueß, C., Obermaier, M., & Reinemann, C. (2018). Perceiving threat and feeling responsible. How severity of hate speech, number of bystanders, and prior reactions of others affect bystanders' intention to counterargue against hate speech on Facebook. *SCM Studies in Communication and Media*, *7*(4), 555-579.

Leung, A. N., Wong, N., & Farver, J. M. (2018). You are what you read: the belief systems of cyber-bystanders on social networking sites. *Frontiers in psychology*, *9*, 365.

Macháčková, H., Dedkova, L., Sevcikova, A., & Cerna, A. (2013). Bystanders' support of cyberbullied schoolmates. *Journal of community & applied social psychology, 23*(1), 25-36.

Madden, C., & Loh, J. (2018). Workplace cyberbullying and bystander helping behaviour. *The International Journal of Human Resource Management*, 1-25.

Munger, K. (2017). Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior, 39*(3), 629-649.

Nesdale, D., Maass, A., Durkin, K., & Griffiths, J. (2005). Group norms, threat, and children's racial prejudice. *Child Development*, *76*(3), 652-663.

Ng, E. (2020). No Grand Pronouncements Here...: Reflections on Cancel Culture and Digital Media Participation. *Television & New Media*, *21*(6), 621-627.

Nocentini, A., Calmaestra, J., Schultze-Krumbholz, A., Scheithauer, H., Ortega, R., & Menesini, E. (2010). Cyberbullying: Labels, behaviours and definition in three European countries. *Australian Journal of Guidance and Counselling*, *20*(2), 129.

Nook, E. C., Ong, D. C., Morelli, S. A., Mitchell, J. P., & Zaki, J. (2016). Prosocial *conformity*: Prosocial norms generalize across behavior and empathy. *Personality and Social Psychology Bulletin*, *42*(8), 1045-1062.

Norris, P. (2020). Closed Minds? Is a 'Cancel Culture'Stifling Academic Freedom and Intellectual Debate in Political Science?.

Obermaier, M., Fawzi, N., & Koch, T. (2016). Bystanding or standing by? How the number of bystanders affects the intention to intervene in cyberbullying. *New media & society*, *18*(8), 1491-1507.

Olweus, D. (1998). Conductas de acoso y amenaza entre escolares [Bullying at school. What we know and what we can do]. Madrid: Morata.

Ozalp, S., Williams, M. L., Burnap, P., Liu, H., & Mostafa, M. (2020). Antisemitism on Twitter: Collective efficacy and the role of community organisations in challenging online hate speech. *Social Media+ Society*, *6*(2).

Paterson, J. L., Brown, R., & Walters, M. A. (2019). The short and longer term impacts of hate crimes experienced directly, indirectly, and through the media. *Personality and social psychology bulletin*, *45*(7), 994-1010.

Patterson, L. J., Allan, A., & Cross, D. (2017). Adolescent perceptions of bystanders' responses to cyberbullying. *New media & society*, *19*(3), 366-383.

Peluchette, J. V., Karl, K., Wood, C., & Williams, J. (2015). Cyberbullying victimization: Do victims' personality and risky social network behaviors contribute to the problem?. *Computers in Human Behavior*, *52*, 424-435.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2020). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 1-47.

Price, D., Green, D., Spears, B., Scrimgeour, M., Barnes, A., Geer, R., & Johnson, B. (2014). A qualitative exploration of cyber-bystanders and moral engagement. *Journal of Psychologists and Counsellors in Schools*, *24*(1), 1-17.

Rafferty, R., & Vander Ven, T. (2014). "I hate everything about you": A qualitative examination of cyberbullying and on-line aggression in a college sample. *Deviant behavior*, *35*(5), 364-377.

Ranieri, M., & Fabbro, F. (2016). Understanding and representing diversity, a media literacy education response to discrimination in news media representations. In J. P. K.Singh, &

T. Hamburger (Eds.).*Media and information literacy: Reinforcing human rights, countering radicalization and extremism* (pp. 109–115). Paris: UNESCO

Salmivalli, C. (2010). Bullying and the peer group: A review. *Aggression and violent behavior*, *15*(2), 112-120.

Schacter, H. L., Greenberg, S., & Juvonen, J. (2016). Who's to blame?: The effects of victim disclosure on bystander reactions to cyberbullying. *Computers in Human Behavior*, *57*, 115-121.

Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media. Association for Computational Linguistics*, 1–10. Valencia: Spain.

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.

Ševčíková, A., & Šmahel, D. (2009). Online harassment and cyberbullying in the Czech Republic: Comparison across age groups. *Zeitschrift für Psychologie/Journal of Psychology*, *217*(4), 227.

Shultz, E., Heilman, R., & Hart, K. J. (2014). Cyber-bullying: An exploration of bystander behavior and motivation. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, *8*(4).

Silverman, T., Stewart, C. J., Birdwell, J., & Amanullah, Z. (2016). The impact of counter-narratives. *Institute for Strategic Dialogue*, 1-54.

Smith, P. K., Mahdavi, J., Carvalho, M., Fisher, S., Russell, S., & Tippett, N. (2008). Cyberbullying: Its nature and impact in secondary school pupils. *Journal of child psychology and psychiatry*, *49*(4), 376-385.

Sponholz, L. (2016). Islamophobic hate speech: What is the point of counter-speech? The case of Oriana Fallaci and The Rage and the Pride. *Journal of Muslim Minority Affairs*, *36*(4), 502-522.

Taylor, S. H., DiFranzo, D., Choi, Y. H., Sannon, S., & Bazarova, N. N. (2019). Accountability and Empathy by Design: Encouraging Bystander Intervention to Cyberbullying on Social Media. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1-26.

Thacker, S., & Griffiths, M. D. (2012). An exploratory study of trolling in online video gaming. *International Journal of Cyber Behavior, Psychology and Learning (IJCBPL)*, *2*(4), 17-33.

Thomas, L., Falconer, S., Cross, D., Monks, H., & Brown, D. (2012). Cyberbullying and the Bystander (Report prepared for the Australian Human Rights Commission). Perth, Australia: Child Health Promotion Research Centre, Edith Cowan University. Retrieved from: https://bullying.humanrights.gov.au/sites/default/files/content/bullying/bystanders/bystanders_results_insights_report.pdf

Tontodimamma, A., Nissi, E., Sarra, A., & Fontanella, L. (2020). Thirty years of research into hate speech: topics of interest and their evolution. *Scientometrics*, 1-23.

Tynes, B. M., Hiss, S., Ryan, A. M., & Rose, C. A. (2015). Effects on mental health and motivation among diverse adolescents in the United States. C.M. Rubie-Davis, J.M. Stephens, P. Watson (Eds.), *The Routledge international handbook of social psychology of the classroom* (pp. 112-121), Routledge, New York, NY.

Van Cleemput, K., Vandebosch, H., & Pabian, S. (2014). Personal characteristics and contextual factors that determine “helping,”“joining in,” and “doing nothing” when witnessing cyberbullying. *Aggressive behavior*, *40*(5), 383-396.

Van Cleemput, K., Vandebosch, H., & Pabian, S. (2014). Personal characteristics and contextual factors that determine “Helping,” “Joining In,” and “Doing Nothing” when witnessing cyberbullying. *Aggressive Behavior, 40*, 383-396.

Van Noorden, T. H., Haselager, G. J., Cillessen, A. H., & Bukowski, W. M. (2015). Empathy and involvement in bullying in children and adolescents: A systematic review. *Journal of youth and adolescence*, *44*(3), 637-657.

Van Royen, K., Poels, K., Vandebosch, H., & Adam, P. (2017). “Thinking before posting?” Reducing cyber harassment on social networking sites through a reflective message. *Computers in Human Behavior*, *66*, 345-352.

Van *Royen*, K., Poels, K., Vandebosch, H., & Adam, P. (2017). “Thinking before posting?” Reducing cyber harassment on social networking sites through a reflective message. *Computers in Human Behavior*, *66*, 345-352.

Walker, J. A., & Jeske, D. (2016). Understanding Bystanders' Willingness to Intervene in Traditional and Cyberbullying Scenarios. *International Journal of Cyber Behavior, Psychology and Learning (IJCBPL)*, *6*(2), 22-38.

Waqas, A., Salminen, J., Jung, S. G., Almerekhi, H., & Jansen, B. J. (2019). Mapping online hate: A scientometric analysis on research trends and hotspots in research on online hate. *PloS one*, *14*(9), e0222194.

Weber, M., Koehler, C., & Schnauber-Stockmann, A. (2019). Why should I help you? Man up! Bystanders’ gender stereotypic perceptions of a cyberbullying incident. *Deviant Behavior*, *40*(5), 585-601.

Weber, M., Viehmann, C., Ziegele, M., & Schemer, C. (2020). Online hate does not stay online–How implicit and explicit attitudes mediate the effect of civil negativity and hate in user comments on prosocial behavior. *Computers in Human Behavior*, *104*.

Young, R., Miles, S., & Alhabash, S. (2018). Attacks by Anons: A Content Analysis of Aggressive Posts, Victim Responses, and Bystander Interventions on a Social Media Site. *Social Media+ Society*, *4*(1), 2056305118762444.

Zwillich, B.J., Haffner, H.P., Bunse, E. (2017) No Place for Hate Speech on Facebook? The Bystander Effect and Intervention Behavior on a Social Network Site. Conference Papers - International Communication Association 2017.