This item is the archived peer-reviewed author-version of:

Harmonizing freedom and protection : adolescents' voices on automatic monitoring of social networking sites

Reference:

Van Royen Kathleen, Poels Karolien, Vandebosch Heidi.- Harmonizing freedom and protection : adolescents' voices on automatic monitoring of social networking sites
Children and youth services review - ISSN 0190-7409 - 64(2016), p. 35-41
Full text (Publisher's DOI): <https://doi.org/10.1016/J.CHILDYOUTH.2016.02.024>
To cite this reference: <https://hdl.handle.net/10067/1321160151162165141>

Harmonizing freedom and protection: Adolescents' voices on automatic monitoring of social networking sites

Van Royen Kathleen, Poels Karolien, Vandebosch Heidi

Abstract

Automatic monitoring of user-generated content on social networking sites (SNSs) aims at detecting potential harm for adolescents by means of text and image mining techniques and subsequent actions by the providers (e.g. blocking users, legal action). Evidently, current research is primarily focused on its technological development. However, involving adolescents' voices regarding the desirability of this monitoring is important; particularly because automatic monitoring might invade adolescents' privacy and freedom, and consequently evoke reactance. In this study, fourteen focus groups were conducted with adolescents (N=66) between 12 and 18 years old. The goal was to obtain insights into adolescents' opinions on desirability and priorities for automatically detecting harmful content on SNSs. Opinions reflect the contention between a need for protection online versus the preservation of freedom. Most adolescents in this study are in favor of automatic monitoring for situations they perceive as uncontrollable or that they cannot solve themselves. Clear priorities for detection must be set in order to ensure the privacy and autonomy of adolescents. Moreover, monitoring actions aiming at the prevention of harm are required.

1. Introduction

Adolescents spend a substantial amount of time on social networking sites (SNSs) (Lenhart et al., 2011). On these sites, adolescents are exposed to risks such as cyberbullying and sexual risks which can result in harm (Lenhart et al. 2011; Ybarra & Mitchell, 2008). Automatic monitoring of user-generated content on SNSs aims at detecting potential harm for adolescents by means of text and image mining techniques and subsequent actions by the providers (e.g. blocking users, legal action). Research in this area has been focusing on its technological development (Cano, Fernandez, & Alani, 2014; Dadvar, Trieschnigg, Ordelman, & de Jong, 2013; Lightbody, Bond, Mulvenna, & Bi, 2014; Van Hee et al., 2015). However, involving adolescents' voices regarding the desirability of this monitoring is important, particularly because automatic monitoring might invade adolescents' privacy and freedom, and consequently evoke reactance (Brehm, 1966). Therefore, in this study we involve adolescents in order to obtain insights into opinions on desirability for automatically detecting harmful content on SNSs.

2. Theoretical background

2.1. Social networking sites: between opportunities and risks

SNSs provide adolescents with several opportunities to actively participate in public culture and to gather and connect with friends (boyd, 2014). In addition, SNSs are important venues for adolescents to experiment with their identity and manage their privacy and intimacy (Livingstone & Brake, 2010). Throughout these activities, youth create new forms of expression and develop social norms in negotiation with their peers (Ito et al., 2008). Several of these created social norms involve the use practises ranging from gossip, flirting, arguing, joking, ostracizing and name-calling (Marwick & boyd, 2011). New ways of communication, experimental forms of self-display and affordances of SNSs, may transform the ample opportunities such as identity performance, intimacy and sociability easily into risks (Livingstone, 2008). Adolescents engage in risk-taking behaviours such as the disclosure of personal information and experimental peer communication, which may exacerbate online risks (Livingstone & Brake, 2010). The EU Kids Online network classified online risks to

children into three types: content risks (content on the web, in which the child is the recipient), contact risks (in which the child is a participant, e.g. being groomed, being bullied) and conduct risks (the child is the actor in creating risks e.g. bullying/harassing) (Hasebrink, Livingstone, Haddon, & Olafsson, 2009). Commonly encountered risks for adolescents on SNSs include cyberbullying and sexual risks (including sexual solicitations and cyber grooming¹) (Lenhart et al. 2011; Ybarra & Mitchell, 2008). However, experiencing risk does not imply harm since harm occurs depending on certain risk factors and protective factors (Staksrud, Ólafsson, & Livingstone, 2013). This study focuses on adolescents' opinions regarding the protection against harm of cyberbullying and sexual risks on SNSs.

2.2. The role of SNSs in providing safety on their platform

In addition to digital literacy and public policy initiatives, one way to protect children is through industry-oriented measures such as monitoring these risks by automatically reviewing user-generated content on SNSs. Industry has been allocated a role in minimizing harm for youngsters on their platforms (Coyne & Gountsidou, 2013). Moreover, the challenge for parents to monitor their children's increasingly mobile media use beyond their view (Haddon & Vincent, 2014) suggests a more prominent role for public policy and industry players in protecting children. Online safety for young users is an important action undertaken by the Digital Agenda of the European Commission and aims at fostering multi-stakeholders dialogue and the self-regulation of SNS providers active in Europe (European Commission, 2010). At the European level, the self-regulating 'Safer SocialNetworking Principles' were formulated, encouraging SNS providers to ensure young users' safety (EC, 2009). These principles serve as a guideline and each provider can apply these recommendations in function of the nature of their network. Principles indicate several options for SNS providers to ensure safety and include providing: clear educational messages and user policies to allow users to navigate their services safely; age-appropriate services (e.g. delete under-age users and promote parental controls); tools and technologies to assist youngsters in managing risks (e.g. privacy settings); easy-to use reporting mechanisms to alarm content violating the Terms of Service; responses to illegal content or conduct; settings for and information on privacy; and assessments of the available services to identify potential risks (EC, 2009). At present, SNSs review their content to detect illegal or prohibited user-generated content, using human moderators or semi-automated forms of detection (Staksrud & Lobe, 2010). However, to keep

track of the vast daily user-generated content, automatic detection systems are suggested (Pachenko, Beaufort, Naets, & Fairon, 2013; Qi & Han, 2007; Reynolds, Kontostathis, & Edwards, 2011). Such systems apply automatic text- and image-categorization techniques using machine learning (Delort, Arunasalam, & Paris, 2011). These are similar to existing effective applications such as spam filtering (Sebastiani, 2002) and can detect potentially harmful content to inform the human moderators of SNSs (who conduct an in-depth analysis of these cases). In addition to words and emoticons expressing insults, profanity and typical contextual words, machine learning models have the capacity to automatically determine gender, age and personality (Schwartz et al., 2013). After detection, SNS providers can further deal with the detected content by taking actions such as removing the content, blocking users, legal action, and offering support. Currently, various efforts are being taken to optimize the detection of content that might represent risks or (associated) mental health problems of adolescents, such as cyber grooming (Bogdanova, Rosso, & Solorio, 2012; McGhee et al., 2011; Peersman, 2012) and cyberbullying (Dadvar et al., 2013; Dinakar, Jones, Havasi, Lieberman, & Picard, 2012; Kontostathis, Edwards, & Leatherman, 2010; Ptaszynski et al., 2010; Van Hee et al., 2015). Since automatic monitoring can take many forms, opinions on desirability will differ depending on what is being detected, who is being monitored and what happens after detection.

In addition to tackling practical and legal difficulties for the role of industry in online protection (Coyne & Gountsidou, 2013) as well as fostering the technological feasibility of automatic monitoring, insights into opinions on the monitoring of SNSs are vital.

2.3 Children's digital rights

This study adopts a child's rights-based approach. The United Nations Convention on the Rights of the Child (UNCRC) (United Nations, 1989) conveys the child's right to be protected against abuse and neglect, urging in online contexts for protection against harm encounters (Third, Bellerose, Dawkins, Keltie, & Pihl, 2014). In particular, beyond protection rights, the UNCRC places equal emphasis on a child's right to participate and the right to freedom of expression (United Nations, 1989). This enshrines the idea to leave children free in seeking, receiving and imparting information online (Livingstone & O'Neill, 2014) and thus, automatic monitoring should embrace this right. Moreover, the automatic monitoring of online interactions may invade youngsters' privacy (van der Zwaan, Dignum, Jonker, & van der Hof, 2014) and consequently, evoke reactance. In advertising, for instance, privacy invasions can

lead to reactance (Brehm, 1966; Tucker, 2014). Nevertheless, privacy is vital to adolescents' lives for developmental goals, including the achievement of autonomy, development of identity, exploration of sexuality and creation of intimacy (Peter & Valkenburg, 2011). This need to respect the child's privacy is asserted by Article 16 of the UNCRC (United Nations, 1989). Respect for privacy and freedom of speech was also argued for by experts in the field of cyberbullying when questioned on the automatic detection of cyberbullying (Van Royen, Poels, Daelemans, & Vandebosch, 2014). Similarly, van der Zwaan et al.'s (2014) framework of desired characteristics for the effectiveness of technologies against cyberbullying (e.g. monitoring, filtering, or educational technology) embodies ethical aspects. Particularly in regard with automatic monitoring technologies, they note that the users' privacy and voluntary use might be at stake (van der Zwaan et al., 2014). Thus, given these ethical aspects, discussions whether to increase the protection of youth online must include children and adolescents' voices to assure their rights (Third et al., 2014). To date, adolescents' perceptions of the desirability of protective strategies to be applied by the industry remain an understudied area, with few exceptions (Byrne & Lee, 2011). Thus far, adolescents have been consulted occasionally regarding their digital rights (Nordic Youth Forum, 2012; Third et al., 2014), parental mediation (Livingstone & Bober, 2003; Media Awareness Network, 2004) or their perceived usefulness of mediation sources (Tomkova, 2012).

In addition to insights into the desirability of automatic monitoring, it is important to investigate what adolescents consider harmful content on SNSs. Not all risks encountered in the online environment result in harm (Staksrud et al., 2013), and it is argued that eliminating all risks is neither feasible nor desirable (Livingstone & O'Neill, 2014). Moreover, focusing on the most harmful situations appears important—as Rooney (2010) suggests, surveillance technologies can negatively impact children by creating a risk-free environment that does not reflect the real world. The 'resilience to risk' approach argues that children can only develop resilience through exposure to risk or stress (Coleman & Hagell, 2007). Children need freedom to experiment and explore the online world, make mistakes and learn to cope, and eventually develop towards resilient individuals (Livingstone & O'Neill, 2014). Therefore, insights into the perceived severity of SNS risks can be useful in this regard.

To summarize, the objectives of this study are to solicit adolescents' views regarding the automatic monitoring of harmful conduct on SNSs and which content should (not) be monitored.

3. Methodology

3.1. Focus group protocol

Focus groups were conducted to obtain data on adolescents' perceptions on desirability of automatic monitoring and perceived harmful content on SNSs. In total, there were 14 focus groups with adolescents aged 12–18 years ($N = 66$) and the group sizes ranged from three to six individuals.

The adolescents were sampled from seven schools in Flanders (Belgium) by means of convenience sampling at the school level. First, the school principal was contacted by mail or telephone and briefed on the purpose of the study through a face-to-face meeting. After obtaining verbal consent from the school principal, adolescents with an active account on an SNS were voluntarily recruited from classes to participate. The researchers briefly explained the study's objectives in each classroom and the adolescents were asked to indicate whether they were willing to participate. For each class, maximum 6 pupils were recruited. Parents of the participants received an information letter, explaining to them the aim of the study, and were given the option to object to their child's participation.

Prior to each focus group, adolescents were informed of the purpose, conditions and potential consequences of the study. All recruited participants provided their written consent to participate and none of their parents objected to their child's participation in the study.

The focus group protocol incorporated theoretical and practical guidelines for focus groups with children (Kennedy, Kools, & Krueger, 2001; Krueger & Casey, 2000). The focus groups comprised as homogenous groups based on gender, education level (vocational or general) and age group (for each grade). Adolescents from grade one were aged 12–14 years old, grade two were 14–16 years old and grade three were 16–18 years old. The focus groups were conducted during October–November 2013.

This study protocol was approved by the Ethical Advisory Committee from the University of Antwerp.

Participants received a gift voucher valued at 10 euros as an incentive to participate. They completed a socio-demographic questionnaire prior to the focus group interview. Audio-

recorded focus groups were led by a trained interviewer. Throughout the interview process, a field journal was maintained to reflect on our own behaviour and background at a later time (Krefting, 1991). A semi-structured questioning route was developed, pre-tested and refined. The issues addressed were designed to solicit information on the desirability, desired characteristics and priorities for the automatic monitoring of harmful content on SNSs. Open-ended questions were followed by more specific probes to clarify and extend responses. Throughout the interview process, questions were repeated or reframed several times (Krefting, 1991). Adolescents were primarily asked what they consider to be severe situations they encountered or could encounter on SNSs and narrate these situations from their own experience as victims or bystanders or heard from peers. Adolescents were asked, following each potential severe case they raised, what they would prefer to happen in this situation and whether SNS providers should exercise more control over these situations. Second, prior to the discussions, the participants were explained the functions of an automatic monitoring system and consequently, were asked for their opinions on this system. Automatic monitoring was neutrally explained as a 'system on SNSs that can automatically detect critical situations through the screening of words and images and consequently, alarms the moderator'.

3.2. Data coding and analysis

A data-driven codebook was created and verbatim transcripts were coded for emerging themes, using a thematic analysis (Boyatzis, 1998). The identified themes related to opinions regarding the desirability of automatic monitoring of adolescents, such as the disadvantages and advantages of monitoring. Second, several codes could be placed under the theme 'desired characteristics'. Adolescents also indicated that certain situations should receive priority for detection, which was assigned to a separate theme named 'priorities'. Summary reports were written according to the identified factors and themes. To increase the rigour of the analysis method, a code–recode procedure was conducted during the analysis phase of the study. Two weeks after creating a codebook by coding a segment of data, the same data were recoded to compare these codes (Krefting, 1991). NVivo software (QSR international — version 10.0) was used to code, manage and analyse the data. In analysing the data, concepts were considered important on the basis of their frequency, specificity, emotion and extensiveness (Krueger & Casey, 2000).

4. Results

4.1. Opinions on automatic monitoring

4.1.1. “Such bad things must be detected and punished”

A general need for higher safety on SNSs was argued for by participants and the majority considers SNSs as important actors in exercising more control. Most adolescents in our sample do not perceive the need for monitoring in all situations (see further in the Section 4.3') and alluded to specific conditions (see further in the Section 4.2) under which monitoring should operate. Adolescents mentioned to prefer to solve the incident themselves by deleting the content or reacting to the perpetrator. When unable to solve or stop the incident themselves, indicated by many as a lack of control, the automatic monitoring system can play a role, according to them. Such situations, in which several adolescents perceive a lack of control, occur, for example, when content cannot be deleted, the perpetrator is anonymous, or there is high public visibility.

The benefits of monitoring SNSs, as cited by adolescents, include ‘serving as a preventive mechanism for harmful behaviour’ and ‘ensuring a safer environment’. Moreover, if monitoring systems would allow faster detection and deletion of content than the current abuse reporting system, adolescents would favour the automatic monitoring. To this effect, a girl (aged 14 years) said the following:

If, for instance, a [harmful] picture is posted, then even before the girl can be insulted through reactions, the picture could be deleted within a minute. That would be very good.

Other positive outcomes attributed to the monitoring system are ‘the identification of the perpetrator in anonymous cases’, the ‘possibility of retributions for the perpetrators’ instead of leaving them unsanctioned and ‘warning perpetrators’ who are crossing the line.

I personally think this [automatic monitoring] would be better. Because everybody nowadays creates a page with bad things and such things must be detected and punished (boy aged 16 years).

4.1.2. *"It is up to the victim to decide to do something about it"*

On the other hand, some adolescents believe it is not the responsibility of social media providers to monitor user-generated content. They consider SNSs to be mere providers of an environment and are not held responsible for the content generated by users.

I compare Facebook with, for instance, this school. There are also people being bullied. And okay, a school can't control this either, but eventually, the people who bully are responsible, whereas a school is merely the place where people gather (girl aged 16–17 years).

A minority of adolescents in this study hold a negative opinion towards automatic monitoring under all circumstances. They fear the loss of autonomy and therefore at first will try to solve the situation themselves by reacting in person.

I think that especially, from where the victim cannot endure it anymore, it is up to the person to decide to do something about it (boy aged 15 years).

In the second place, they prefer to report harmful content to the provider themselves if necessary. However, adolescents argued for a fast response on these 'abuse' reports, as many frustrations exist regarding the inefficiency of the reporting system. Some adolescents, regardless of opinion, are concerned about their 'privacy'. These concerns relate in particular to the detection of sexual risks, since they linked it to the monitoring of private messages. However, overall, we observed that privacy was mentioned to a limited extent.

But maybe it is a bit awkward, if you, for instance, don't have a problem and they detect something. They penetrate into your private life, among your friends, and they can read what you write (boy aged 13 years).

A few adolescents mentioned they do not want to give up their excitement on SNSs, as they like to be ‘bystanders’ and see (when they are not the victim) the occurrence of something sensational like cyberbullying or the sharing of nude pictures.

The point is, I think, if you see that and you don't know the person, you may think ‘wow yes, how does she look like!’ [about someone depicted as a whore]... Well, why do people watch Jersey Shore? That is also [sensational] (girl aged 17 years).

Another risk of automatically monitoring content on SNSs, alluded to by adolescents, is the loss of freedom of expression and spontaneity.

Then maybe SNSs would become too decent, and then many people would not like it [these platforms] anymore. If they have the feeling, that each reaction they write, can be read by someone they don't know, then they would not like it anymore (boy aged 13 years).

4.1.3. Doubts on feasibility

Irrespective of their opinions, adolescents have many doubts related to monitoring, mainly due to the thin line between joking and bullying.

This concern of ‘potential false labelling of perpetrators’ is a major reason for the lack of support for monitoring. Moreover, adolescents emphasized that each individual has personal boundaries, which renders it impossible for outsiders to judge whether someone perceives a certain situation as harmful.

I also think that SNSs don't know whether ‘it is a joke or not’,...and I think, if they are always going to remove that [the content], I think many people won't be interested anymore in SNSs, or at least youth won't. Because such a joke is common for us. As long as it is a joke, it is not bad. Okay, if it is not meant as a joke but SNS providers don't know when... when it is a joke or not (boy aged 15 years).

Another reason for the lack of faith in detection, as adolescents suggested, is the presence of numerous people and events on SNSs. Some adolescents believe in the possibility to circumvent the monitoring system, for instance by using other words or a pseudonym. Adolescents raised concerns regarding a reactive approach of monitoring. Some stated that an SNS cannot solve critical incidents and if detected, this would mean that the message had already been delivered and caused harm.

4.2. Requirements for automatic monitoring

Adolescents, even those who are not in favour of automatic monitoring, mentioned several characteristics for the automatic monitoring system to become desirable.

4.2.1. Involve the perceived victim

To address the concern of differences in individual boundaries and the difficulty in recognizing cyberbullying, they thought it is necessary to contact the perceived victim in case of detection and ask him or her whether the content is actually considered as harmful.

If you are being contacted, then I think it can be a positive idea. As long as it is not every day that you get such a question like ‘do you think this can stay online?’ But then you can say yourself as well ‘I don't want it’, then it is positive, otherwise...(boy aged 15 years).

Involving the perceived ‘victim’ is important, as some of them would not consider themselves as such. For instance, some respondents mentioned that sometimes girls do not mind to be on pages like ‘sexy teenagers’. Moreover, it was mentioned that an agreement should be obtained as to whether the help of parents or other adults can be involved, since they are believed to overreact and escalate the issue. The escalation of an event such as cyberbullying or the sharing of nude pictures is a concern for adolescents and a reason to solve it independently. Adolescents stressed the confidentiality of detected content and that this should not be passed to third parties without consent. It was mentioned several times that asking consent in general is necessary, particularly when screening private messages.

4.2.2. No exaggeration

Related to concerns of losing freedom, adolescents require that an automatic detection system does not detect all types of incidents and safeguard the autonomy of adolescents in dealing with these events.

4.2.3. Preventive monitoring

As previously mentioned, for some situations, the content should be prevented from being dispersed, including nude pictures, public hate pages and fake profiles.

Just, they have to check it, the people from Facebook, they have to check it first and if they see that it is not appropriate for being publicly visible, they should not allow it to be uploaded (boy aged 14 years).

Adolescents believe that a prohibition of anonymous comments and acts on SNSs would refrain (some) people from bullying and increase control on the side of the victim. Moreover, a certain identity control should be conducted to avoid the creation of fake accounts, such as logging in with an identity card, similar to bank accounts. Another suggestion for prevention is warning those who engage in risky behaviour, such as maintaining a publicly visible profile or adding/accepting too many unknown friend requests.

4.2.4. Follow-up is crucial

In addition, a response is required after detection, such as actions taken against the perpetrator. For adolescents, deleting the content is not considered sufficient. Suggested responses towards the perpetrator include sending warnings or blocking (temporarily) his or her account. This is important to prevent perpetrators from engaging in the behaviour again.

It's about trying to change his opinion. Because by only removing that reaction, he's going to react again to other pictures. You have to confront because only removing has no use (girl aged 16–17 years).

Follow-up actions such as helping and supporting the victim were only rarely mentioned. However, in general, in addition to automatic monitoring, participants require more practical information on how to cope with bullying on SNSs such as how to report it and block a person as well as the provision of direct support links.

It is especially important that people know how to prevent and that they know what they can do. Because when I was being bullied on Facebook, I didn't know what I could do, I didn't know how I have to report someone or block someone (girl aged 16–17 years).

4.3. Priorities for monitoring

Adolescents specified situations that should be priority as well as what should not be detected. They disagree on whether to monitor cyberbullying and harassment. The monitoring of bullying evokes many doubts owing to the feasibility to detect it and because autonomy is at stake. Comments and tags that can be removed by the victim require no monitoring.

But if it is a picture that you can remove yourself, then it is ok. But from the moment you can't remove it yourself, then I think they can detect it (girl aged 14–15 years).

Almost all adolescents agree that the following forms of cyberbullying should be given priority in detection: public hate pages and fake profiles that damage the reputation of victims (for instance, through impersonation). Hate pages refer to a page created on Facebook about someone that contains derogatory comments. They are perceived as threatening because of their public character (visible to all users), the use of pictures without consent, the anonymity of perpetrators (no possibility to approach the perpetrator) and the impossibility to remove them yourself. These characteristics can be considered in determining further priorities. Second, fake profiles refer to stealing someone's identity. Their impact is high according to adolescents because people can do things in someone else's name to damage his/her reputation, either as a perpetrator of bullying or by approaching someone for sexual purposes. On the other hand, one can also be approached by someone with a fake profile for purposes of sexual solicitation, which increases the lack of control. Adolescents expect from automatic detection the prevention of a fake profile use or retrieval of the perpetrator.

All situations in which (real or manipulated) nude pictures are involved are considered severe. Given this concern, a strong need is expressed for the prevention of uploading, the inability of copying pictures and the block of the use of someone's pictures without consent.

Importantly, feelings of helplessness are experienced if everyone has already seen or shared them. In addition to nude pictures, other situations in which videos and pictures of someone are used without consent should be monitored.

Other aggravating aspects we could deduce from severe bullying or harassment stories include: higher numbers of people reacting negatively to a situation, multiple perpetrators and sexualized cyberbullying focusing on looks or sex life.

Sexual solicitations are not considered severe because of their perceived high controllability. Controllability, in this case, refers to the ability to easily block the person on an SNS, ignore the event and remove the content as solicitations are most often sent through private messages. Continued sexual requests in which coercion is used by adults receive support for detection, not because the victim cannot escape or ignore it but rather to catch the perpetrator. A girl aged 15 years said,

I would like the person to be traced... if it is a paedophile, gosh!

Not only girls but also younger boys are in favour of monitoring cyber grooming. Older adolescents prefer protection only for younger adolescents and stressed one's own responsibility and control in this regard.

5. Discussion

This study employed a qualitative methodology to investigate adolescents' opinions on the monitoring of harmful situations on SNSs. The findings suggest that most adolescents in this study are not negatively inclined towards the automatic monitoring of user-generated content on SNSs. However, they do not feel the need for being monitored in situations they can solve themselves and believe that specific conditions should be met for monitoring to be desirable. A duality can be observed in their opinions between, on the one hand, concerns related to their loss of freedom, autonomy and privacy and on the other hand, the need for monitoring 'uncontrollable' situations (i.e. those they cannot solve themselves), confirming the importance of balancing both their rights of protection and participation.

5.1. Maintaining autonomy, freedom of expression and privacy while monitoring

If confronted with a threatening situation, adolescents prefer to report or solve the incident independently (e.g. delete the content or react). A minimal interference with the adolescents' autonomy is important as it is a crucial developmental task during adolescence (Silverberg & Gondoli, 1996) and is important to build resilience (Livingstone & O'Neill, 2014). Another reason for the desire to maintain autonomy, indicated by the adolescents, is to avoid escalation.

Previous studies support this point by showing that adolescents fear exacerbation when disclosing cyberbullying to adults (Mishna, Saini, & Solomon, 2009) and do not reveal incidents of traditional bullying because of fear for retaliation (Mishna & Alaggia, 2005).

Second, adolescents raised concerns regarding the loss of freedom of expression, since they use SNSs to experiment with self-expression and communicative norms (Ito et al., 2008) and thus may perceive automatic monitoring by SNS providers as a threat to their interactions in the SNS space. The fear for feeling inhibited to spontaneously interact and the sense of an overprotected environment may provoke a shift of problematic behaviour to other non-monitored environments.

Third, some adolescents mentioned the loss of privacy as a potential drawback of automatic monitoring. However, they were mostly concerned about the involvement of parents. This corresponds with previous findings that SNS users prefer to control their information with respect to their social environment (i.e. social privacy) more than with respect to companies or institutions (i.e. institutional privacy) (boyd, 2014; Raynes-Goldie, 2010). Adolescents develop several strategies to achieve their privacy from adults such as parents on SNSs (boyd, 2014), which explains their fear for parents knowing about critical incidents. Moreover, even though adolescents were in favour of detecting cyber grooming, privacy concerns were issued regarding intimate private messages. However, it is vital to safeguard adolescents' privacy to allow sexual exploration (Peter & Valkenburg, 2011), as SNSs provide adolescents with new venues for intimacy practices alike flirting and dating beyond the gaze of adults (Pascoe, 2009).

5.2. Where to draw the line?

Additionally, the adolescents raised several practical and ethical questions with automatic monitoring. Important to note is the complexity of drawing the line between appropriate and harmful content. Many doubts were indicated regarding the feasibility of automatic monitoring and the associated concern of ‘false labelling’. A tendency exists in which the forms of interaction are too easily labelled as ‘bullying’ (Guldberg, 2009). For instance, forms of interaction as drama practises or teenage conflict are often recognized as bullying by adults (Marwick & Boyd, 2011). The detection of irony and sarcasm also poses a challenge for natural language processing techniques to automatically detect cyberbullying (Filatova, 2012; Reyes, Rosso, & Buscaldi, 2012). Moreover, the line between victims and perpetrators of bullying is often difficult to draw as they may coincide (Mascheroni & Olafsson, 2014). Similarly, for cyber grooming and sexting the line between victims and offenders is often blurred (van der Hof & Koops, 2011). Innocent conversations may falsely being labelled as sexual harassment or may seem harmful to adults but not perceived as such by the minor. Also sexting (i.e. the sending and receiving of sexually suggestive or explicit pictures through electronic means) stirred much debate whether to consider it normal intimate communication within romantic and sexual relationships or to criminalize these practises (Angelides, 2013; Simpson, 2013).

5.3. Monitoring between protection and freedom

To provide an answer to these questions and needs of adolescents, we propose to set clear priorities for detection. These include severe situations such as in which a loss of control is perceived by many adolescents, including mostly public hate pages, fake accounts set up to harass them and (nude) pictures shared without consent. A core aspect in rating the severity of a situation was the extent of control, referring to public and anonymous acts (Sticca & Perren, 2012). Anonymous acts are perceived as less solvable as they impede the victim to react, and highly visible public content owing to the loss of control and incapability of deleting content intensify negative interactions (boyd, 2009).

Next, incorporating measures such as consulting the perceived victim after detection, are important. In addition, alternative automatic monitoring strategies respecting the freedom of expression may be the use of ‘reflective interfaces’ (Dinakar et al., 2012; Jones, 2012). This includes notifications such as warnings that urge users to reflect in anticipation of or following to posting potentially harmful content on SNSs, but do not restrict such behaviours (Dinakar et al., 2012). Other preventive options that were preferred was the filtering of nude

pictures and severe public hate pages before being publicized. Also providing effective responses towards perpetrators to impede future perpetration was highly favoured, as well as better responses by the providers to reported content by adolescents themselves. Finally, more monitoring options for adolescents in order to allow to remove victimizing content by themselves, may also be a good strategy to guarantee their autonomy.

Another question concerns the age of the group to be monitored. Detection of grooming must be limited to certain age categories, as for older children, privacy and/or freedom of expression might outweigh the need for protection (van der Zwaan et al., 2014). Younger adolescents in this study felt a higher need to monitor sexual risks, suggesting a targeted approach with focus on protecting younger adolescents. Studies demonstrate that younger adolescents are more upset by sexual risks (Livingstone & Görzig, 2014; Mitchell, Finkelhor, & Wolak, 2001; Ringrose, Gill, Livingstone, & Harvey, 2012).

Finally, to counteract reactance and to comply with European data protection laws (Council of the European Union, 1995), transparent information on the procedure of automatic monitoring would be important and have to be adapted to youth in a comprehensible way.

5.4. Limitations and future research

This study contributes to the ongoing debate on the degree to which the youth can be protected and are willing to be monitored on SNSs. Even though different opinions were observed within the same focus group, the method of interviewing adolescents in small groups may have incited peer pressure and socially desirable answers for some adolescents. As such the use of focus groups may impose a limitation to the findings. Moreover, these results may not be generalizable to other populations. Adolescents' opinions must be critically reviewed. The arguments of some adolescents contained a strong element of moral disengagement (Bandura, Barbaranelli, Caprara, & Pastorelli, 1996), attributing the responsibility of harassment to the victim. This practise of victim blaming and diffusing personal responsibility has been previously associated with cyber victimization (Pornari & Wood, 2010). Moreover, the findings did not reveal clear patterns of personal differences (for age and gender) regarding the desirability for protection. Throughout the focus groups adolescents may have expressed desirability for monitoring their vulnerable peers, whereas they may not perceive monitoring to be necessary for themselves. If we would involve more actual victims, it may be likely that we would find a higher desirability for monitoring among

girls, as they are generally more upset by contact risks (Vandoninck, d'Haenens, & Roe, 2013).

Further research is needed on how to monitor and protect vulnerable adolescents. Moreover, actual victims should be involved to know their expectations with regard to follow-up. In addition, the views of health professionals, educators and parents are required.

Cross-cultural research should be conducted, as disagreement may exist in the acceptability of monitoring. Cultural differences have been observed to the effect of online privacy perceptions (Soffer & Cohen, 2015).

Furthermore, the high perceived severity of pictures being misused argues for more research on cross-media detection techniques (i.e. text and image detection) (Lightbody et al., 2014), in addition to studies focusing on the textual detection of risks (Dinakar, Reichart, & Lieberman, 2011; Kontosthatis et al., 2010).

6. Conclusions

Measures to ensure prevention of harm on SNSs, such as filtering pictures with nudity before being uploaded, warning about potentially harmful behaviour and providing effective responses towards perpetrators to impede future perpetration, are highly favoured. These actions must be further examined with respect to effectiveness and to ensure increased accurate detection. Second, most desirable situations to be monitored by SNS providers are those that victims cannot solve themselves. To meet adolescents' concerns, protection must be harmonized with safeguarding freedom by, for instance, consulting the perceived victim and setting clear priorities. We argue to further open up the debate on how we can protect children and adolescents on social media and involve them more in this debate.

Acknowledgements

This work was supported by the Agency for Innovation by Science and Technology (IWT) under Grant 120007.

References

- Angelides, S. (2013). "Technology, hormones, and stupidity": The affective politics of teenage sexting. *Sexualities*, 16(5/6), 665–689.
- Bandura, A., Barbaranelli, C., Caprara, G., & Pastorelli, C. (1996). Mechanisms of moral disengagement in the exercise of moral agency. *Journal of Personality and Social Psychology*, 71(2), 364–374. <http://dx.doi.org/10.1037/0022-3514.71.2.364>.
- Bogdanova, D., Rosso, P., & Solorio, T. (2012). On the impact of sentiment and emotion based features in detecting online sexual predators. *Proceedings of the 3rd workshop on computational approaches to subjectivity and sentiment analysis* (pp. 110–118). Jeju: Republic of Korea.
- Boyatzis, R. E. (1998). *Transforming qualitative information: Thematic analysis and code development*. London: SAGE Publications.
- boyd, d. (2009). *Friendship. Hanging out, messing around, geeking out: Living and learning with new media* (pp. 79–115). Cambridge: MIT Press.
- boyd, d. (2014). *It's complicated: The social lives of networked teens*. New Haven, CT: Yale University Press.
- Brehm, K. (1966). *A theory of psychological reactance*. Oxford, England: Academic Press, 135.
- Byrne, S., & Lee, T. (2011). Toward predicting youth resistance to internet risk prevention strategies. *Journal of Broadcasting & Electronic Media*, 55(1), 90–113. <http://dx.doi.org/10.1080/08838151.2011.546255>.
- Cano, A., Fernandez, M., & Alani, H. (2014). *Detecting child grooming behaviour patterns on social media*. SociInfo 2014. Barcelona, Spain.
- Coleman, J., & Hagell, A. (2007). *Adolescence, risk and resilience: Against the odds*. West Sussex: John Wiley & Sons Ltd.
- Council of the European Union (1995). Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, Pub. L. No. OJ L 281.
- Coyne, I., & Gountsidou, V. (2013). The role of industry in reducing cyberbullying. In P. K. Smith, & G. Steffgen (Eds.), *Cyberbullying: Research on coping with negative and enhancing positive uses of new technologies* (pp. 83–98). London: Psychology Press.
- Dadvar, M., Trieschnigg, D., Ordelman, R., & de Jong, F. (2013). Improving cyberbullying detection with user context. *Proceedings of the 35th European conference on IR research*, vol. 7814. (pp. 693–696). Berlin: Springer-Verlag Berlin, Heidelberg.
- Delort, J. -Y., Arunasalam, B., & Paris, C. (2011). Automatic moderation of online discussion sites. *International Journal of Electronic Commerce*, 15(3), 9–30. <http://dx.doi.org/10.2753/JEC1086-4415150302>.
- Dinakar, K., Reichart, R., & Lieberman, H. (2011). Modeling the detection of textual cyberbullying. presented at the international AAAI conference on weblogs and social media, Barcelona, Spain. Retrieved from <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/3841/4384>

- Dinakar, K., Jones, B., Havasi, C., Lieberman, H., & Picard, R. (2012). Commonsense reasoning for detection, prevention and mitigation of cyberbullying. *ACM transactions on interactive intelligent systems*, 2(3). <http://dx.doi.org/10.1145/2362394.2362400>.
- EC Social Networking Task Force (2009). Safer social networking principles for the EU. Luxembourg: European Commission (Retrieved from https://ec.europa.eu/digitalagenda/sites/digital-agenda/files/sn_principles.pdf).
- European Commission (2010). Digital agenda for Europe. Action 37: Foster self-regulation in the use of online services. Retrieved from <http://ec.europa.eu/digital-agenda/en/pillar-iii-trust-security/action-37-foster-self-regulation-use-online-services>
- Filatova, E. (2012). Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, & S. Piperidis (Eds.), *Proceedings of the eight international conference on language resources and evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA).
- Guldborg, H. (2009). Reclaiming childhood. Freedom and play in an age of fear. Vol. The Bullying Bandwagon Abingdon, Oxon: Routledge, 92–110.
- Haddon, L., & Vincent, J. (2014). European children and their carers' understanding of use, risks and safety issues relating to convergent mobile media. (No. Report D4.1). Milano: Unicatt (Retrieved from http://eprints.lse.ac.uk/60147/1/_lse.ac.uk_storage_LIBRARY_Secondary_libfile_shared_repository_Content_Net%20Children%20Go%20Mobile%20Project_Reports_European%20children%20and%20their%20carers%E2%80%99%20understanding_2014.pdf).
- Hasebrink, U., Livingstone, S., Haddon, L., & Olafsson, K. (2009). Comparing children's online opportunities and risks across Europe: Cross-national comparisons for EU Kids Online (2nd ed.). LSE, London: EU Kids Online (Retrieved from http://eprints.lse.ac.uk/24368/1/D3.2_Report-Cross_national_comparisons-2nd-edition.pdf).
- Ito, M., Horst, H., Bittani, M., Boyd, D., Stephenson, B., Lange, P., & Robinson, L. (2008). Living and learning with new media: Summary of findings from the digital youth project (p.58). Chicago, IL: The MacArthur Foundation.
- Jones, B. K. (2012). Reflective interfaces : Assisting teens with stressful situations online (unpublished PhD dissertation). Massachusetts institute Of technology. Retrieved from <http://dspace.mit.edu/handle/1721.1/76519>.
- Kennedy, C., Kools, S., & Krueger, R. (2001). Methodological considerations in children's focus groups. *Nursing Research*, 50(3), 184–187. <http://dx.doi.org/10.1097/NNR>.
- Kontosthatis, A., Edwards, L., & Leatherman, A. (2010). Text mining and cybercrime. *Text mining: Applications and theory* (pp. 149–163). West Sussex: John Wiley & Sons Ltd.
- Krefting, L. (1991). Rigor in qualitative research: The assessment of trustworthiness. *The American Journal of Occupational Therapy*, 45(3), 214–222. <http://dx.doi.org/10.5014/ajot.45.3.214>.
- Krueger, R., & Casey, M. (2000). Focus groups: A practical guide for applied research (3rd ed.). London: Sage Publications, Inc.
- Lenhart, A., Madden, M., Smith, A., Purcell, K., Zickuhr, K., & Rainie, L. (2011). Teens, kindness and cruelty on social network sites. Pew Internet research Project (Retrieved from <http://www.pewinternet.org/2011/11/09/teens-kindness-and-cruelty-onsocial-network-sites/>).
- Lightbody, G., Bond, R., Mulvenna, M., & Bi, Y. (2014). Investigation into the automated detection of image based cyber bullying on social media platforms. (2 pp.) Presented at the 2014 Irish machine vision and image processing conference. Derry-Londonderry

Northern Ireland: IPRCS.

- Livingstone, S. (2008). Taking risky opportunities in youthful content creation: teenagers' use of social networking sites for intimacy, privacy and self-expression. *New Media & Society*, 10(3), 459–477.
- Livingstone, S., & Bober, M. (2003). UK children go online : Listening to young people's experiences. London: LSE Research Online.
- Livingstone, S., & Brake, D. (2010). On the rapid rise of social networking sites: New findings and policy implications. *Children & Society*, 24(1), 75–83. <http://dx.doi.org/10.1111/j.1099-0860.2009.00243.x>.
- Livingstone, S., & Görzig, A. (2014). When adolescents receive sexual messages on the internet: Explaining experiences of risk and harm. *Computers in Human Behavior*, 33, 8–15. <http://dx.doi.org/10.1016/j.chb.2013.12.021>.
- Livingstone, S., & O'Neill, B. (2014). Chapter 2 Children's rights online: Challenges, dilemmas and emerging directions. *Minding minors wandering the Web: Regulating online child safety* (pp. 19–38). Berlin: Springer.
- Marwick, A., & boyd, d. (2011). The Drama! Teen conflict, gossip, and bullying in networked publics (SSRN scholarly paper No. ID 1926349). Rochester, NY: Social Science Research Network.
- Mascheroni, G., & Olafsson, K. (2014). Net children go mobile. Risks and opportunities (2nd ed.). Milano: Educatt.
- McGhee, I., Bayzick, J., Kontosthatis, A., Edwards, L., McBride, A., & Jakubowski, E. (2011). Learning to identify internet sexual predation. *International Journal of Electronic Commerce*, 15(3), 103–122. <http://dx.doi.org/10.2753/JEC1086-4415150305>.
- Media Awareness Network (2004). Young Canadians in a wired world: Phase II. Focus groups. Ota: Media Awareness Network.
- Mishna, F., & Alaggia, R. (2005). Weighing the risks: A child's decision to disclose peer victimization. *Children & Schools*, 27(4), 217–226. <http://dx.doi.org/10.1093/cs/27.4.217>.
- Mishna, F., Saini, M., & Solomon, S. (2009). Ongoing and online: Children and youth's perceptions of cyber bullying. *Children and Youth Services Review*, 31(12), 1222–1228.
- Mitchell, K., Finkelhor, D., & Wolak, J. (2001). Risk factors for and impact of online sexual solicitation of youth. *JAMA*, 285, 3011–3014. <http://dx.doi.org/10.1001/jama.285.23.3011>.
- Nordic Youth Forum (2012). Youth have their say in internet governance. Stockholm: The International Clearinghouse on Children, Youth and Media (Retrieved from http://www.nordicom.gu.se/sites/default/files/publikationer-hela-pdf/youth_have_their_say_web.pdf).
- Pachenko, A., Beaufort, R., Naets, H., & Fairon, C. (2013). Presented at the ECIR'13 proceedings of the 35th European conference on advances in information retrieval, Heidelberg. Towards detection of child sexual abuse media: Categorization of the associated filenames, vol. 7814. (pp. 776–779), 776–779.
- Pascoe, C. (2009). Intimacy. Hanging out, messing around, geeking out: Living and learning with new media (pp. 117–148). Cambridge: MIT Press.
- Peersman, C. (2012). Conversation level constraints on pedophile detection in chat rooms. CLEF 2012 conference and labs of the evaluation forum (pp. 1–13).
- Peter, J., & Valkenburg, V. (2011). Adolescents' online privacy: Toward a developmental perspective. In S. Trepte, & L. Reinecke (Eds.), *Privacy online: Perspectives on privacy and self-disclosure in the social web* (pp. 221–234). Heidelberg: Springer.

- Pornari, C. D., & Wood, J. (2010). Peer and cyber aggression in secondary school students: The role of moral disengagement, hostile attribution bias, and outcome expectancies. *Aggressive Behavior*, 36(2), 81–94. <http://dx.doi.org/10.1002/ab.20336>.
- Ptaszynski, M., Dybala, P., Matsuba, T., Masui, F., Rzepka, R., Araki, K., & Momoushi, Y. (2010). In the service Of online order: Tackling cyber-bullying with machine learning and affect analysis. *International journal of computational linguistics research*, 1, 135–154.
- Qi, X., & Han, Y. (2007). Incorporating multiple SVMs for automatic image annotation. *Pattern Recognition*, 40(2), 728–741. <http://dx.doi.org/10.1016/j.patcog.2006.04.042>.
- Raynes-Goldie, K. (2010). Aliases, creeping, and wall cleaning: Understanding privacy in the age of Facebook. *First Monday*, 15(1) (Retrieved from <http://firstmonday.org/ojs/index.php/fm/article/view/2775/2432>).
- Reyes, A., Rosso, P., & Buscaldi, D. (2012). From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering*, 74, 1–12. <http://dx.doi.org/10.1016/j.datak.2012.02.005>.
- Reynolds, K., Kontostathis, A., & Edwards, L. (2011). Using machine learning to detect cyberbullying. *Proceedings of the 2011 10th international conference on machine learning and applications and workshops — Volume 02* (pp. 241–244). Washington, DC, USA: IEEE Computer Society. <http://dx.doi.org/10.1109/ICMLA.2011.152>.
- Ringrose, J., Gill, R., Livingstone, S., & Harvey, L. (2012). A qualitative study of children, young people and “sexting.” NSPCC. Retrieved from http://www.nspcc.org.uk/inform/resourcesforprofessionals/sexualabuse/sexting-research-report_wdf89269.pdf
- Rooney, T. (2010). Trusting children: How do surveillance technologies alter a child's experience of trust, risk and responsibility? *Surveillance & Society*, 7(3/4), 344–355.
- Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., ... Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS One*, 8(9), e73791. <http://dx.doi.org/10.1371/journal.pone.0073791>.
- Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1), 1–47. <http://dx.doi.org/10.1145/505282.505283>.
- Silverberg, S., & Gondoli, D. (1996). Autonomy in adolescence: A contextualized perspective. *Psychosocial development during adolescence: Progress in developmental context* (pp. 12–61). Thousand Oaks, CA: SAGE.
- Simpson, B. (2013). Challenging childhood, challenging children: Children's rights and sexting. *Sexualities*, 16(5–6), 690–709. <http://dx.doi.org/10.1177/1363460713487467>.
- Soffer, T., & Cohen, A. (2015). Privacy perception of adolescents in a digital world. *Bulletin of Science, Technology & Society*, 0270467615578408. <http://dx.doi.org/10.1177/0270467615578408>.
- Staksrud, E., & Lobe, B. (2010). Evaluation of the implementation of the safer social networking principles for the EU part I: General report. Luxembourg: European Commission Safer Internet Programme.
- Staksrud, E., Ólafsson, K., & Livingstone, S. (2013). Does the use of social networking sites increase children's risk of harm? *Computers in Human Behavior*, 29, 40–50.
- Sticca, F., & Perren, S. (2012). Is cyberbullying worse than traditional bullying? Examining the differential roles of medium, publicity, and anonymity for the perceived severity of bullying. *Journal of Youth and Adolescence*. <http://dx.doi.org/10.1007/s10964-012-9867-3>.

- Third, A., Bellerose, D., Dawkins, U., Keltie, E., & Pihl, K. (2014). Children's rights in the digital age. A download from children around the world. Melbourne: Young And Well, 92.
- Tomkova, J. (2012). Mediation of safer internet use. Bratislava: Research Institute for Child Psychology and Pathopsychology (Retrieved from http://www.zodpovedne.sk/download/Prieskum_Mediacia_EN.pdf).
- Tucker, C. E. (2014). Social networks, personalized advertising, and privacy controls. *Journal of Marketing Research*, 51(5), 546–562. <http://dx.doi.org/10.1509/jmr.10.0355>.
- United Nations. (1989, November 20). Convention on the rights of the child. Retrieved from <http://www.ohchr.org/Documents/ProfessionalInterest/crc.pdf>.
- van der Hof, S., & Koops, B. -J. (2011). Adolescents and cybercrime: Navigating between freedom and control. *Policy & Internet*, 3(2), 1–28. <http://dx.doi.org/10.2202/1944-2866.1121>.
- van der Zwaan, J. M., Dignum, V., Jonker, C. M., & van der Hof, S. (2014). On technology against cyberbullying. In B. van den Berg, & B. Schermer (Eds.), *Minding minors wandering the Web: Regulating online child safety* (pp. 211–228). The Hague: Asser Press.
- Van Hee, C., Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., De Pauw, G., ... Hoste, V. (2015). Detection and fine-grained classification of cyberbullying events. *Proceedings of the 10th recent advances in natural language processing (RANLP 2015)* (pp. 672–680).
- Van Royen, K., Poels, K., Daelemans, W., & Vandebosch, H. (2014). Automatic monitoring of cyberbullying on social networking sites: From technological feasibility to desirability. *Telematics and Informatics*, 32(1), 89–97. <http://dx.doi.org/10.1016/j.tele.2014.04.002>.
- Vandoninck, S., d'Haenens, L., & Roe, K. (2013). Online risks: Coping strategies of less resilient children and teenagers across Europe. *Journal of Children and Media*, 7(1), 60–78. <http://dx.doi.org/10.1080/17482798.2012.739780>.
- Ybarra, & Mitchell, K. (2008). How risky are social networking sites? A comparison of places online where youth sexual solicitation and harassment occurs. *Pediatrics*, 121(2), e350–e357. <http://dx.doi.org/10.1542/peds.2007-0693>