Tobe: A Virtual Keyboard and an Animated Character for Individual and Educational Cyberbullying Intervention

NOA MORAG YAAR, Media Innovation Lab Reichman University Israel, Israel
EDEN GROSSMAN, Media Innovation Lab Reichman University Israel, Israel
SAGI HATAN, Media Innovation Lab Reichman University Israel, Israel
NOAM KIMCHI, Media Innovation Lab Reichman University Israel, Israel
OFIR NASH, Media Innovation Lab Reichman University Israel, Israel
HADAS EREL, Media Innovation Lab Reichman University Israel, Israel

As the use of communication on social media among children increases, the cyberbullying phenomena is becoming a prominent challenge. We present the preliminary design of Tobe, a virtual keyboard that provides textual and visual feedback in real-time to help prevent harmful discourse on social media among elementary school-aged children. The widget is designed based on two unique principles: (1) An empathy based feedback mechanism which involves an animated character and verbal statements (both positive and negative); (2) A class specific solution in which children and the teacher build a vocabulary of words together. This element provides transparency as to why feedback was shown at any given moment, and can increase cooperation and empowerment for both the teacher and the children. We present the human centered design process and early results from a user study with 12 children. Our findings show high engagement with the system, and user awareness before texting.

CCS Concepts: • Human-centered computing → Human computer interaction (HCI): HCI design and evaluation methods;

Additional Key Words and Phrases: Cyberbullying, Virtual keyboard, Reflective technologies, Empathy, Real-time Feedback, Elementary school

ACM Reference Format:

1 INTRODUCTION
Cyberbullying among children is an urgent concern with profound implications for emotional health and well-being [Olweus 2012]. It is a broad phenomenon that may lead to emotional distress with extreme impact on self-esteem and psychological resilience [Hamm et al. 2015; Tokunaga 2010]. It is considered one of the most harmful types of violence that exists among school-aged children [Aizenkot 2018]. These consequences are most significant in late elementary school years as it is a critical period for the development of mental health [Kowalski et al. 2014]. With digital apps and social networks becoming an integral part of the space in which children and adolescents conduct social connections, cyberbullying becomes even a greater concern [Marzano 2020] due to a rise in hate messages, harassment, threats, and exclusion from social groups [Marzano 2020; Mishna et al. 2010]. Cyberbullying is defined as intentional aggressive behavior, perpetrated by a group or individual, using electronic means of contact, against a victim who cannot easily defend himself/herself [Smith et al. 2006]. Victims typically report negative psychological consequences, including emotional distress, depression, stress, anxiety, feelings of loneliness, sleep disturbance and drug usage [Perren et al. 2010]. In extreme cases, cyberbullying can lead to various psychopathologies such depression, anxiety, and suicidal thoughts [Kowalski et al. 2014]. Interestingly, there are also indications for profound negative impact on the offenders. Cyber-offenders commonly experience mental and psychological distress including depression and anxiety, as well as various social difficulties [Cross et al. 2015].

There are common technical strategies for dealing with cyberbullying, such as blocking the offender or deleting the text messages. Other methods include deleting anonymous messages without reading them, and reporting the user [Smith et al. 2006]. Though these strategies can be helpful in some cases, it requires the victim to be proactive in the situation they are usually ashamed of, or prevents them from seeking help from relevant adults [Cassidy et al. 2009; Dehue 2013; Slonje et al. 2013; Smith et al. 2006]. To try to deal with that, some educational intervention programs are conducted in schools with the aim to increase awareness, and encourage empathy [Hutson et al. 2018].

There are also technological solutions that were specifically designed for cyberbullying. These typically involve widgets designed to either block bullying messages, or to provide information on solutions and coping strategies in response to cyberbullying [Slonje et al. 2013]. While these solutions can assist in preventing specific cyberbullying occurrences or provide help and support for victims, they are not designed for behavior change and raising children’s awareness of the consequences of their actions. Such an approach that addresses the offenders instead of providing a technical solution or support for the victims, may lead to a profound change that will reduce cyberbullying in general [Cross et al. 2015]. An example for technology that takes such an approach is “ReThink”, a keyboard extension that detects offensive words and warns users to stop and think again before sending an offensive message. The aim of the keyboard is to prevent cyberbullying and lead to general behavior change [Campbell and Bauman 2018].
In this work, we explore the possibility of extending the idea of a virtual keyboard by designing a solution that is based on reflection to evoke empathy and suggesting it to be class specific. In our solution, feedback, both negative and positive, is enhanced by an animated character. Hence, similar to "ReThink", the keyboard monitors the child’s text. However, apart from identifying offensive words, it also identifies positive words and concertises its consequences by impacting the character’s state. The animated character is designed to provide reinforcement when the child chooses positive words, and alerts when the child chooses offensive text. In addition, the solution is designed to be a part of an educational program, where the children and the teacher are defining the positive and offensive words together thus creating a personalized class vocabulary. While conversing on social media, Tobe collects anonymously the words from the vocabulary database that were used by children when texting to their classmates. Both the construction of the class’s vocabulary and the anonymous data, are allowing the teacher to be involved and to initiate an informative discussion in class.

2 RELATED WORK
Most technologies designed for dealing with cyberbullying focus on automatic identification of cyberbullying text, followed by notifying relevant adults (parents, teachers, etc) or blocking of offensive text. For example “BullyBlocker” is an app that identifies instances of cyberbullying in Facebook and notifies parents when it occurs [Silva et al. 2016]. Another example is “BullStop”, a mobile app for detecting and preventing cyberbullying on social media platforms. The app uses deep learning models for identifying offensive text and automatically initiates actions such as deleting offensive messages and blocking offenders [Salawu et al. 2020].

A somewhat different approach involves technologies for preventing cyberbullying by increasing empathy and raising awareness of consequences. These technologies are typically based on Reflective User Interfaces. These interfaces are designed for encouraging positive digital behavioral norms by leading to reflection on behavior in real-time and allowing the users to choose whether to send the offensive message or not [Dinakar et al. 2012]. The association between the real-time reflection and the following immediate action is assumed to mitigate offensive comments and enhance positive digital behavior. Reflective User Interfaces include notifications, action delays, and displaying consequences. By warning users that they are about to send offensive text, they provide them with an opportunity to stop and think again [Campbell and Bauman 2018].

In this work, we present a preliminary design of a system that couples opportunities raised by both approaches. The virtual keyboard is designed as a Reflective User Interface that enhances empathy using an animated character. The textual feedback and the change of mood of the animated character increases engagement in the reflection process. The keyboard preserves the child’s control and responsibility and does not force any decision, increasing the chances that the child will accept it as an add-on tool. In addition, the keyboard is designed to be an integral part of a supervised educational program in the children’s school. Within the class context, children build their own vocabulary with the teacher (both positive and negative), and the data is analyzed and sent to the teacher anonymously for follow-up activities in the classrooms.

3 DESIGN
The design process follows the Human Centered Design approach (HCD) which takes into consideration values, pains and motivation of the stakeholders. HCD key principles of validation of need and solution with the end users through user study in an iterative process, was at the center of our design.

3.1 Need Study
To identify the need, we started with preliminary interviews with relevant stakeholders: an online crime investigator and school psychologist. We realized that cyberbullying has increased among school students because of the COVID-19 [Investigator]. The education system has failed to produce stability and response to students in real time. The crime investigator said that “children have spent more time on social networks and the number of reported vulnerabilities has increased significantly from 2019 to 2020”. Their input on the fact that cyberbullying occurs mostly after school hours was supported by the work of scholars in the field [Wollack and Mitchell 2000]. Moreover we sent a short questionnaire to parents of children aged 10-18 (n = 55). The questionnaire included 10 questions about cyberbullying, whether their child has been bullied in the past, have they known about it and from whom, how they dealt with it once they realized their child was bullied, etc. Our findings show that parents are typically unaware or informed after the incident. More than 50% of parents reported that their children rarely share bullying incidents with them. Some also periodically check their children’s phones. Another finding was that children do not always understand that they have been bullied.

![Fig. 1. An empathy map of our end-user to understand her needs in a specific context. The map contains relevant quotes and data we collected, ordered by how the user thinks and feels, what she sees, hears and do, in a specific context](image-url)
3.2 Design guidelines based on need study and literature

With these insights we conducted in-depth semi-structured interviews with 5 experts: middle school counselor, sixth grade educator, boarding school coordinator, and went back and interviewed the school psychologist and the online crime investigator once more. In addition we interviewed 7 parents of preschool children. Using a thematic coding analysis [Boyatzis 1998; Gibbs 2008], in search for repeating themes in the data, while allowing new themes to emerge [Fereday and Muir-Cochrane 2006; Ghosh et al. 2018], several insights were derived: 1) Cyberbullying varies consistently between digital platforms, but significant assaults occur in the WhatsApp application. This finding was also found in a study among fourth- to twelfth-grade students, where 31% of the students reported being victims of cyberbullying in WhatsApp groups (including insults, cursing, group removal and posting offensive images [AIZENKOT 2018].

2) Children do not share their online experiences, 'Adolescence is all 'black or white', they do not see parents as an address to contact' [School Psychologist]; 'children would prefer to share with their peers or outsiders before contacting their parents' [P5]. 3) There is high parental stress due to the need to identify and deal with their children’s cyberbullying, 'I have an obsessive anxiety about what my child is feeling in such situations' [P3]. On one hand they feel they have to invade their children’s privacy to seek information and on the other hand they are worried that it will threaten their relationship. Their anxiety arises also in their lack of ability to deal with assaults when they appear, 'I am completely losing control, I do not know what is right and wrong and how to act' [P6].

4) Bullying behavior is typically impulsive and expresses a lack of empathy. This might have been avoided if a deeper thinking process was applied: 'In the heat of the moment people do not pay attention and therefore do not stop themselves'. [P1]

Based on these insights we defined a persona. We conducted additional interviews with 5 children aged 8-14 (n=5) and created an empathy map to pinpoint the user’s feelings, thoughts, and behavior at a given moment (See figure 1). The process resulted in a persona between the ages 8-9 (4th grade), an age where children begin to communicate online. Following are the design guidelines that were derived based on the map.

Following the literature review and the need study we derived three design principles: 1) **Real-time intervention.** The fact that cyberbullying occurs in digital platforms offers the opportunity for real-time intervention. This intervention can lead the user to reconsider or give her/him an opportunity to not send the offensive message. 2) **Reflect to evoke empathy.** Unlike face-to-face bullying, cyberbullying is hidden in the way the impact of the negative behavior has on the victim. Therefore the chances of empathy and remorse are significantly reduced [Kowalski et al. 2014]. Providing opportunity for reflection that is known to increase empathy 3) **The context of the class.** A solution should be intertwined with a class activity in addition to providing children with the opportunity to avoid negative messaging by reflection. For greater effectiveness in a non-intrusive way, the solution should involve an educational figure. The context of the class is indicated as having the greatest potential for relevant intervention, if the cyberbullying reports are anonymous and only allow for general class interventions by the teacher.

These design principles have led us to the metaphor of a door stop. A door stop prevents the door from slamming and shutting the entrance, or from hitting against the wall. It helps the door become resilient to sudden wind. Similarly, the purpose of the keyboard is to prevent children from "hitting" their friends on social media, insults that can occur abruptly and lead to drastic consequences. In addition, the door stop is not fixed and can be removed or adjusted, so does our widget. The children with the teacher can modify the vocabulary so it suits the condition/atmosphere of the class, and it can also be removed and children can choose not to follow the feedback, without judgment.

4 THE SYSTEM

Tobe is a virtual keyboard and a friendly character that helps prevent abusive online discourse among elementary school-age children. Tobe works on every social media platform and can recognize offensive and positive words in real time. Tobe reacts to an offensive message by visual and textual feedback just before the message is sent.

After downloading the widget, the children receive a unique code from the teacher. With the completion of a short on-boarding process (name, gender), the widget performs 3 main interactions: 1) Creating a class vocabulary, 2) Presents real-time textual and graphical feedback on any texting platform, 3) Collects data anonymously for class discussions.

In the on-boarding process (see figure 2), children create two vocabularies. One of positive words that they would like to receive, and one of abusive words that they would not like to receive from their classmates. Each child can contribute to the vocabulary anonymously as part of a class activity led by the teacher. Overtime, Tobe will learn dynamically as more words are flagged by the user as hurtful words. It is also possible to manually add words in additional class sessions per the teacher’s decision and guidance. By contributing words to the class vocabulary, children feel responsible and engaged. This will also make the technological system and the feedback mechanism more transparent to them as they understand why a word was flagged and initiated a feedback. The teacher’s responsibility is to allow an open, non judgmental discussion over words presented by the children.
4.1 Real-time Feedback
Following the "reflection to evoke empathy" design principle, we examined a few directions for evoking empathy: a picture, a textual feedback, Emoji and/or a graphical character. We ruled out the idea to evoke empathy by using the children’s pictures since we found that using a picture of the victim and/or the offender can cause distress and tag the children as "bully" or "weak". We decided to present textual feedback and together with a UX writer we created an array of short sentences that would be influential. Two types of feedback were designed: 1) When a child writes a positive word detected from the class vocabulary, she will be praised for doing that, e.g. “You are truly awesome", 2) When detecting negative words, Tobe presents 2 types of textual feedback: a) Gentle textual sentence below the insert line, e.g. "Are you sure you want to send this"? b) A black pop-up which blocks the keyboard and needs to be dismissed before she can continue (Extreme feedback). This type of feedback will only appear if an offensive message was sent by the user more than 3 times.

Complimentary to the textual feedback, we also designed a third feedback as an animated virtual character. This character is presented to the side of the textual feedback (see figure 3). The character responds to the written words by changing its color and appearance. The character either represents sadness (sad face colored in red for offensive words) or happiness (happy face colored in green for positive words) in response to the child's text in real time. We relied on prior research, which found that prosocial cartoon characters can increase children’s empathy levels as well as reduce their level of aggression [Zhang and Cao 2021].

4.2 The System’s Data
The class vocabulary is dynamic in the sense that it adds new words and improves over time. Words can be added manually by the children and the teacher around a class activity. All added words are reviewed by the teacher. In addition, with the consent of the parents, the widget collects data anonymously (words used in the text message, time, quantity) and produces reports for the teacher. (see figure 4). The purpose is to help the teacher tailor educational activities that are relevant to the online experiences. For example, if the teacher understands that the word "gay" is repeated and used abusively, he can carry out an educational activity on the subject of sexual identity.

5 PRELIMINARY EVALUATION
To gain initial insights concerning children’s attitudes towards the virtual keyboard and to evaluate its usability we conducted a preliminary evaluation study under strict COVID-19 safety regulations.

5.1 Participants and Procedure
Initial wire-frames were tested with 12 children, (8 boys, 4 girls, between the ages of 8.5-12). The researcher explained to them the scenario and asked them to complete two tasks: 1) On-boarding and contributing words (positive and negative) to the vocabulary as if they were with the teacher in class, 2) Simulate a conversation between 2 children.

Due to ethics considerations we chose not to simulate a bullying scenario that could lead to discomfort or anxiety. Instead, we asked the children to think out loud and share their thinking process while simulating common texting with a friend. To support the thinking out loud processes we encouraged children to share their thoughts by asking "What would you do?", "What did you expect to happen?", "What do you think the character feels?". After the experience, we conducted a semi-structured interview with the children where they were asked to describe their thoughts and ideas about the keyboard and virtual character. Initial user study towards low fidelity prototype.

Fig. 3. Improving the feedback by adding a "clear" button next to the statement

5.2 Preliminary Findings
The thematic analysis processes resulted in 5 main themes; 1) On-boarding process was clear, participants understood that it was an educational process. At the same time, some had difficulty understanding the meaning of the entry code (so we decided to clarify this point in his sentence "Class Code: Get a code from the teacher). 2) The virtual character was perceived as a positive characteristic of the keyboard and children liked it. They understood the meaning and the color change from green (happy) to red (sad) and the connection between its appearance and the text. 3) Children understood the two types of negative feedback and stated that the extreme feedback was very effective. "It really makes me stop and think about whether to send" (P8, Female). 4) Children missed the positive feedback. To address that we added a 3rd purple color to emphasize a neutral conversation, thus when the character changes from purple to red (sad) or green (happy), it’s association to the text would be clear. 5) Some children suggested adding game play features such as earning stars or points that can facilitate other activities such as personalized design, skin and colors of Tobe. 6) Children understood the "clear" button next to the textual feedback and thought it was more influential.

5.3 Mid Fidelity Prototype
Based on our initial findings we performed an additional iteration that involved changes to the feedback, and the virtual character. We changed the position of the virtual character on the keyboard and changed the order of the screens, allowing a clearer on-boarding experience in which the character is presented before starting to add
words to the vocabulary. In terms of the data, we designed a desktop interface for the teacher. On the screen, the teacher can filter the data by period of time, gender and common words (positive and negative). We also introduced a vocabulary management interface for the teacher with the capability of editing as well as approving new words added by the children. The data shown can assist the teacher with directing her efforts in class, for example if she learns that more girls use negative words than boys. The anonymity of the data is kept in order to maintain the kids’ privacy and control.

Fig. 4. A system generated report which anonymously reflect the class discourse for the teacher only

6 DISCUSSION
In this work we designed and implemented a non-invasive technological solution for promoting healthy behavior on social media. The solution, a virtual keyboard with an animated character that responds to the text’s valence, is designed based on two approaches commonly used for dealing with cyberbullying: 1) using reflection for increasing empathy [Hutson et al. 2018], 2) involving adults that can assist in mediating such complex and sensitive situations [Salau et al. 2020; Silva et al. 2016]. The system provides real-time feedback when children send the message, verifying that the child reflects on the text before sending it and reducing impulsive behavior. The feedback is both positive and negative, aiming to increase positive behaviors among the children without judgment. An anonymous summary of children’s behavior and texting is sent to the teacher that can design relevant educational activities accordingly.

The feedback mechanism that we created is designed to be non-intrusive and aims to increase empathy by allowing textual and visual feedback which is both positive and negative. The virtual character that children connect with is designed to enhance empathy and increase engagement in positive rather than negative texting. The direct influence of the chosen wording and the character’s state is designed to concretize the possible consequences of the text message. In the preliminary evaluation, children connected with the character and believed that the reflection processes will be highly effective. Specifically, the positive reinforcement is supported by the “social learning theory”, according to which positive reinforcement increases the likelihood that a welcomed behavior will occur more frequently [Maisto et al. 1999]. In addition, positive reinforcements have been found to help build a child’s self-image. While cyberbullying often results in insecurity and distress, positive feedback has the potential to lessen its negative impact. In addition, creating gentle feedback that does not block the user’s actions, provides the child with a sense of control that will prevent children from rejecting the app.

Giving the widget the context of the class and allowing the teacher to be part of the solution is another key feature in our design. The process of creating the vocabulary together reinforces the responsibility children have for each other. It also makes the system more transparent to the children as they understand why the feedback was shown to them. The potential to provide information about the state of the class and raise the teachers’ awareness to dangerous situations allows for relevant interventions led by professional educators which can eventually reduce cyberbullying events in the class and provide help if an event has occurred.

7 LIMITATIONS AND FUTURE DIRECTION
This study has some limitations. First, the product has not been tested in an educational setting with children and with an educational figure and therefore further research is needed. To measure the effectiveness of Tobe we propose to conduct a long term study in the educational frameworks that will examine the implications and effects of using this product, and whether the violent discourse is inhibited or diminished. The study should also identify the most appropriate feedback type and frequency that would lead to a balance between preserving children’s control while raising their awareness. Second, ideally we would like Tobe to be installed on children’s smartphones, however we understand that regulatory processes will prevent the installment of a built-in keyboard. Third, even though our product collects data anonymously, we need the parent’s consent to be able to access this data and by that keep the children’s privacy. Forth, future studies should also evaluate the possibility of including additional stakeholders such as parents. Finally, we acknowledge that personality traits vary between children and can lead to different effects.

8 CONCLUSION
To conclude, we present a virtual keyboard enhanced by an animated character to be used among elementary school children who share the same class. The combination of adding a reflection process with concertizes visualization of a character to evoke empathy among children, is only to be emphasized by placing the solution within a school class setting. Both stakeholders are engaged, and the teacher has the potential to ameliorate one of the most urgent problems on social media by creating a pleasant social environment for the children.

ACKNOWLEDGMENTS
We wish to thank the following people for their valuable help and advice: Ofir Sadka, Eran Ilani, Andrey Grishko, Amit Dekalo.
REFERENCES


