Developing strategies to stop bullying: Design considerations for an artificial training center

Sylvie Saget
University of Gothenburg
sylvie.saget@gu.se

Alina Yarmina
Moscow State University
a.yarmina@gmail.com

Marina Escobar Planas
Joint Research Centre
European Commission
marescplajob@gmail.com

Abstract

Reporting incidents to an adult is the top recommendation for youth victims facing bullying. At the same time, victims need to develop strategies to use when they are facing their offenders: counter aggression and making a safety plan. This paper presents design considerations for a conversational system being a training center for a victim to learn and try such strategies. We first detail what is bullying and existing preventive measures. We then detail specific features of such a conversational system and we define a set of functionalities and dialog design considerations required to ensure compliance with the preceding features. We close with the presentation of a simple illustrative implementation.

1 Introduction

School violence and bullying occurs throughout the world and affects a significant proportion of children and teenagers. It is estimated that 246 million children and teenagers experience school violence and bullying in some form every year (UNESCO, 2017). Results of the meta-analysis of 80 research papers devoted to the problem of bullying among 12-18 aged teenagers (both aggressors and victim roles were analyzed) show that the average level of bullying frequency is about 35% (Modecki et al., 2014), whilst in the other research data it varies from 9% to 98%. The issue also has great importance as susceptibility to systemic school bullying directly affects physical and mental health, general well-being and academic performance of young people (Pells et al., 2016). Studies show that bullying is strongly associated with social maladjustment, including vandalism (Solberg and Olweus, 2003), fights (Nansel et al., 2003), drug usage and smoking (Hinduja and Patchin, 2008; Shetgiri, 2013; Hemphill et al., 2012), truancies at school (Byrne, 1994; Shetgiri, 2013), weapon carrying (Nansel et al., 2003; Shetgiri, 2013), and other form of antisocial behavior (Solberg and Olweus, 2003). It is also proved that bullying situation negatively influence on each of its participants.

This paper presents design considerations for a conversational system being a training center for a victim to learn and try strategies to use while facing their offenders. We first detail what is bullying and existing preventive measures. We then detail specific features of such a conversational system as being a tool to embed in an anti-bullying program, the need of transparency when it comes to defining the expected relationship between the system and the victim, as well as safety issues. In the second section, we also define a set of functionalities and dialog design considerations required to ensure compliance with the preceding features. We close with the presentation of a simple illustrative implementation.

2 Bullying

2.1 What is bullying?

Generally, bullying is being defined as intimidation, humiliation, harassment, physical or psychological terror. Aggressive behavior can be defined as bullying when it becomes prolonged and systematic, aimed at the person unable to protect him/herself in these circumstances. Bullying is a form of aggressive behaviour behaviour designed to hurt another. There is not universal agreement on the definition of bullying, but there is some consensus that it is aggressive behaviour which satisfies two additional criteria: (1) repetition it happens more than once and (2) there is a power im-

balance such that it is difficult for the victimto defend himself or herself (Olweus, 1999). Among main types of bullying are: 1) direct bullying: physical (kicks, blows), verbal (insults, threats coming up with unpleasant nicknames), damaging property and personal belongings 2) indirect: the spread of false rumors and gossip, exception from social groups. Researchers also distinguish several roles which an teenager can assign to him/herself in any bullying situation: these are chameleon, observer, aggressor, and victim.

School bullying is widespread (not to say universal), and because of that, it seems to be studied well. It is being studied all over the world, and there is a constant growth of publications on the topic during the last 30 years. Still, more complicated forms and types of bullying appear. Besides the traditional forms of bullying, in the modern digital world, it becomes easier to bully someone using ICTs (Information and Communication Technologies) so that we can observe the emergence of cyberbullying. Thus, it becomes very hard to get fully protected from all bullying forms, so each person under 18 can enter to the at-risk zone as a potential victim.

In general, bullying can be described as a global and very traumatic but still not fully studied phenomenon. Types of bullying evolve with time, so it is essential to develop preventive measures and methods of struggle against it using the latest technologies and instruments.

3 Robotheraphy

3.1 International and national anti-bullying-programs

Experience of European research teams shows that preventive measures can be realized at different levels, from local to national, and international. The first national program has been developed in 1938 by D. Olweus and called the Olweus Bullying Prevention Program (OBPP). It is based on several fundamental principles: friendly positive attitude, concerted applying of sanctions in case of unacceptable behavior/violation of rules, adults involvement as a role model (Olweus and Limber, 2010). Since 2001, the Olweus program has a status of priority all-national project, which is been being widely used in Scandinavian countries, Austria, Germany, and Iceland. The next anti-bullying project KIVA (kiusaamistavastaan - against humiliations) (Garandeau et al., 2014). It has become a national program, and since 2009 has been being implemented in 90% Finnish schools, and several EU countries. Among the main blocks of the KIVA program are prevention, methodical instruments for influencing bullying behavior in situations where real violent acts emerge, regular at-school monitoring. Both programs are complex and include both counteraction and prevention.

There also several anti-bullying social competency national programs which have particular importance as they are primarily aimed to upgrade children prosocial skills. For instance, in United Kingdom, SEBS (Social, Emotional and Behavioral Skills) program has been designed to develop 3-11 aged preschool/primary school attendants social, emotional and behavioral skills, and is being financed by the State Education Department. Program participants upgrade self-awareness, emotional control, motivational regulation, empathy, etc. The other program Second step has been founded at the USA (Seattle, Washington state Committee on children) in 1978. It includes several general educational strategies: discussions, storytelling, master classes, and activities aimed to develop and strengthen children individual qualities. Finally, in 1996, Roots of Empathy program for preschool and school attendants (up to the 8-th grade of secondary school) has been elaborated in Canada (Toronto). It includes the following blocks: emotional competency, acceptance, involvement, attachments, violence prevention, etc.

At the local level, each educational institution can choose the program or design a plan of anti-bullying events autonomously. The UNESCO Bureau offers the following general guidelines for educational staff: monitoring and analysis of the actual situation, developing common at school anti-bullying policy, event planning and management, psychological service work organization, education and informational events for all participants of educational process teachers, parents and children, design of informational materials. Nowadays academics and educators find more and more effective methods to reduce the bullying frequency. E.g., Trofi and Farington (Ttofi and Farrington, 2011) analyzed 44 high-quality school-based intervention programs and found that on average, these reduced bullying by 20-23% and victimization by 17-20%. Unfortunately, there is still no universal educational program able to fully eliminate the bullying rate. It also has to

be said that nearly all existing anti-bullying programs have to be updated because the majority of them were designed 20-25 years ago, and do not consider the evolvement of bullying types. Classic anti-bullying programs are gradually replaced by the new ones, where the usage of ICTs (including machine learning) is provided. Interdisciplinary research teams design special applications and software and even use robots (Fear Not! (Watson et al., 2007), Media heroes (Chaux et al., 2016), etc.). in order to reduce both offline and online bullying frequency.

3.2 Therapy using human-robot interaction

Recent advances in robotics have also enabled social robots to fulfill a variety of functions in the psychotherapeutic process. A social robot may be defined as an artificially intelligent system that has a physical embodiment, is autonomous or semi-autonomous, and interacts and communicates with humans by following the behavioral norms expected by the people with whom the robot is intended to interact (Bartneck and Forlizzi, 2004).

In the paradigm developed by Feil-Seifer and Mataric (Feil-Seifer and Mataric, 2005, p. 465), the robots goal is to create a close and effective interaction with a human user for the purpose of giving assistance and achieving measurable progress in convalescence, rehabilitation, learning, and so forth. Libin and Libin (Libin and Libin, 2004, p. 1792-1793) also tried to define the role of the robot in human-robot interactions and they introduced the term robotherapy, defined as a framework of human-robotic creature interactions aimed at the reconstruction of a persons negative experiences through the development of coping strategies, mediated by technological tools, to provide a platform for building new positive life skills. David, Matu, and David (David et al., 2014, p. 4) suggested that the term robotherapy should be replaced by robot-assisted/enhanced therapy and defined as the use of robots in a personalized evidence-based psychotherapy framework, where the robot should be seen as a technological tool that can help the psychotherapists to accomplish their clinical roles and aims.

3.3 Anti-bullying program using human-robot interaction

It the article, we present a working scheme with the teenager who has faced up with the traditional face-to-face bullying. The specificity of the model is that it involves individual work with a child, and the uniqueness is that it can be used in any place where the robot is present. To make the model work, a specialist does not need to involve any pedagogic or administrative school staff. The target auditory is teenagers who are currently being bullied directly (a victim who regularly faces up physical/verbal aggression). The scheme includes psychologist-teenager and teenager-robot interacting. It has 3 basic stages:

1. **Psychologist-teenager interaction.** In this session, a psychologist clarifies the psychological issue. The teenager is being bullied face-to-face and is unable to protect him/herself. After working with the situation and resources search, a specialist offers to a teenager to train his response to the aggressor.

2. **Teenager-robot interaction.** A psychologist helps the teenager to switch on the robot and leaves the room. The robot asks for the teenager agreement to get involved in a robotherapy and informs teenager about the start of the session (introduction). Finally, a bullying scenario takes place.

3. **Psychologist-teenager interaction.** Specialist enters the room and asks about the robot-interaction session results. Also, psychologist evaluates the emotional condition and mood of the client. It is also assumed that the session with a psychologist will be held later so that the teenager could assimilate the experience.

The robot can also be used by the educational staff for reducing the bullying rate. In case there is an teenager in the group who can not effectively react to aggressors, he/she can be proposed to try the robot interaction which will improve his/her skills to give the adequate responses to aggressors, and improve self-confidence.

4 Design considerations

In this section, we detail our three design principles: safety, distributed decision, and transparency regarding robot's function, as well as the corresponding understanding abilities.

4.1 Safety

While existing systems such as FearNot (Watson et al., 2007) allow children to experiment strate-

gies without being directly involved, the idea behind our solution is to allow teenagers to directly experiment such strategies. This kind of system is especially challenging regarding safety issues as teenager will have to face aggressive behaviours from the system.
According to Hawkins & al. "When peers intervene against bullying, 57% of bullying episodes cease within 10 seconds" (Hawkins et al., 2001). Therefore, encouraging such peer intervention is an important approach to reduce bullying. However, despite the effectiveness of peer intervention, Hawkins et al. found that although bystanders were present in 88 % of bullying episodes, they only intervened and defended victims in 19% of cases. Then, empowering victims themselves with the ability to intervene against bullying may be a powerful way to address this issue. To ensure safety of the approach, teenagers has been chosen as they may be more skilled and have a deeper understanding about their safety compared to child. In addition, the system is not supposed to be used "into the wild" but under the control of a therapist, as detailed in the previous section.

However, as pointed out by Bickmore and al. (Bickmore et al., 2018), any conversational agent dedicated to health care has to address numerous challenges: privacy of the information shared with the robot, how the information will be stored and used later, frustration and anger of not being understood. Our design principle is to ensure safety by a "qualitative therapeutic alliance" (Horvath et al., 2011) both in the design of the system itself and by ensuring the coherence of the overall therapeuthic program.

4.2 Distributed decision

Three agents are engaged in a robotherapy: the psychologist, the teenager and the robot. The three of them may decide on parameter or action:

- **Psychologist:** as the leader of therapeutic process, she is the first in charge of teenager's safety. She also needs to provide highly personalize therapy in regards of the complexity of bullying. Moreover, at the end the first step of a therapy, she is able to evaluate teenager's resources and she has to indicate what to train.

- **Robot:** The robot is responsible of the training itself. It must be able to ensure teenager's safety all along a training session. Indeed, the psychologist may observe a training session and may come back in the room whenever she decides. However, the robot may exhibit a coherent behaviour as being part of the therapeutic team. Meaning, it may have a rational behaviour generally speaking and may take responsibility of stopping a training session if the teenager's safety is not guaranteed anymore.

- **Patient:** The teenager needs to be actively engaged in therapy to ensure therapy's effectiveness. The aim is also for him to develop assertive behaviour. Then the more he is engaged in decision and actions, the best it is. In our proposal, it is strictly mandatory that the teenager is the one who takes the decision to start a anti-bullying session.

Key decisions are about:

- being engaged in a robotherapy, when a training session starts/stops;

- which training scenarios and what degree of aggressiveness are acceptable and suitable;

- acceptable and unacceptable strategies.

A distributed decision regarding these questions is notably a key element for a coherent and qualitative therapeutic alliance. If one want to establish a qualitative alliance, three factors may have to be taken into account: (a) the collaborative nature of the relationship, (b) the effective bond between patient and therapist, and (c) the patient's and therapist's ability to agree on treatment goals and tasks (Martin et al., 2000).

4.3 Transparency

The robot being part of a therapeutic alliance, its role, limitations and responsibilities may have to be made explicit and indicated to the teenager in a complementary and possibly repetitive way during the psychologist-teenager interaction and at the beginning of teenager-robot interaction. The patient may have the opportunity to agree/disagree with such a role before starting any training session. This has to be included in the dialog flow as being part of a mandatory introduction.

Moreover, the robot has two "roles" during a training session. It acts as a therapist. Typically, during the introduction, it aims to ensure that the

teenager is aware and agree with therapeutic principles. During the training itself, it tracks possible emotional distress signs and reacts accordingly. It also behaves as an actor simulating to be a bully. We suggest as a design principle robot's transparency regarding which mode the current actions of the robot correspond to. This can be done either explicitly using natural language indicating a change of mode, or by a color code of robot's eyes, for instance.

4.4 Understanding the patient

In order to maintain a user model and to act accordingly, we now need to specify what information the robot is supposed to perceive. The robot observes patient for two reasons: to manage the training session and to stop it if the teenager's safety is not guaranteed anymore.

In order to ensure safety by a qualitative therapeutic alliance, the range of strategies to stop bullying accepted by the system may have to be validated both through a deep understanding of bulling and by the therapist. Actually, the range of counter-aggression is large and includes unacceptable strategies such as the use of weapons (Black et al., 2010). Fighting back is also risky. Then, a special attention has to be given to the specification of acceptable strategies. However, we integrated counter-aggression because studies point it as a common reaction of victims facing bullying (Black et al., 2010). It is also important to point that mindset and regulation/policy act have to change accordingly - as it does not fit with zero-tolerance policies. Counter-aggression may be qualified as bullying notably because bullies know how to turn the system into their advantage.

The association "Eyes on bullying"[1] published a toolkit specifying insights, strategies, activities, and resources to address bullying. Here are their recommendation regarding strategies for standing up to bullies (Storey et al., 2013, p. 9):

- take a deep breath and let the air out slowly;
- sit or stand tall, head up;
- keep your hands at your sides rather than on your hips or folded across your chest;
- have a relaxed and purposeful facial expression, not angry or laughing;
- maintain eye contact;
- speak with a calm voice, loud enough to be heard clearly;
- use non-provocative words and a confident tone of voice;
- avoid name-calling or making threats;
- avoid finger pointing or other threatening gestures;
- reply briefly and directly. Avoid bringing up past grudges or making generalizations (You always...).

Tracking teenager's emotional distress seems more complicated. As far as we know, there is not a similar list of typical behavior in the bullying context. Such a tracking may rely on a multimodal detection mixing complex emotion detection, such as sadness or fearness, and pure physical indicator such as big eyes or static posture.

5 A simple prototype

The anti-bullying program is an idea developed during the Human Robot Interaction project work of the HUMAINT Winter School on Artificial Intelligence and its ethical, legal, social and economic impact which took place February 4-8th 2019[2] at Seville Joint Research Centre. In order to illustrate our idea, we sketch a simple prototype using Choregraph and the robot humanoid Pepper. Choregraph is a program to build behaviours in Softbank robotics robots (Nao and Pepper). A behaviour is how we call a program that allow us to tell the robot what to do. Fig. 1 illustrates the Anti-Bullying Behaviour.

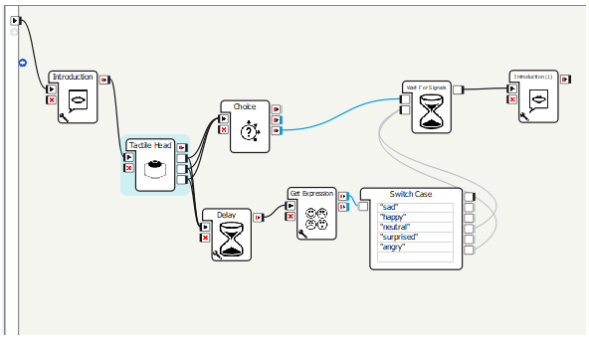

Figure 1: Anti-Bullying Behaviour

[1] http://eyesonbullying.org

[2] https://ec.europa.eu/jrc/communities/en/comment/1039

The program has different parts:

- **Introduction:** The robot welcomes the child, explains the program to him and gives instructions:

  Hello. Welcome to the private antibullying program. Here you will learn how to react when you face an offender. Our conversation is confidential, so please, feel free to say whatever you want. Also remember this is a simulation, and we will stop it whenever you need it. Let's train now. To start a bullying scenario, please touch my head.

  After the speech box, the tactile sensor of the head is activated, allowing the robot to know when the child is ready. Once the output is active, two different boxes are activated. One for bullying simulation and another one for its evaluation.

- **Bullying simulation:** In this module, the robot emulates a verbal aggression with body movement to the kid and waits for the child's reaction.

- **Evaluation:** The robot tries to recognize child's expression and verbal reaction. There are three different options:
  - any error happened during recording,
  - the given answer is not a good one (one that helps the child to face bullying),
  - the given answer is a good one (one that helps the child to face bullying).

- **Congratulation:** If the child gives what we called "good answer", the robot congratulates the child, and apologies for its behavior remembering that it was just a simulation.

6 Conclusion

In this paper, we aimed to enlighten the utility to empower victim of bullying with the ability to defend themselves. Conversational AI is especially interesting in that context as it enables such victims to develop such skill training with a non-human before being ready to face humans. However, such a system is challenging especially when it comes to guarantee patient's safety. Such a system is also challenging for the Conversational AI community as bullying scenario is an understudied type of dialog. Detection of cyberbullying is an emerging research question (Van Hee et al., 2018) but such a system would require even more if we target to go beyond highly scripted bullying scenarios. Turn-taking strategy - including speed - seems to be an interesting aspect as it's related to aggressiveness (Ter Maat et al., 2010).

Acknowledgement

The research reported in this paper was supported by a grant benefited from the support of HUMAINT winter school. We would like to thank Vicky Vasiliki, who organised the "Human-Robot Interaction" project session. This research was also supported by a grant from the Swedish Research Council for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

References

Christoph Bartneck and Jodi Forlizzi. 2004. A design-centred framework for social human-robot interaction. In *RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No. 04TH8759)*, pages 591–594. IEEE.

Timothy Bickmore, Ha Trinh, Reza Asadi, and Stefan Olafsson. 2018. Safety first: Conversational agents for health care. In *Studies in Conversational UX Design*, pages 33–57. Springer.

Sally Black, Dan Weinles, and Ericka Washington. 2010. Victim strategies to stop bullying. *Youth violence and juvenile justice*, 8(2):138–147.

Brendan Byrne. 1994. *Coping with bullying in schools*. Continuum International Publishing Group.

Enrique Chaux, Ana María Velásquez, Anja Schultze-Krumbholz, and Herbert Scheithauer. 2016. Effects of the cyberbullying prevention program media heroes (medienhelden) on traditional bullying. *Aggressive behavior*, 42(2):157–165.

Daniel David, Silviu-Andrei Matu, and Oana Alexandra David. 2014. Robot-based psychotherapy: Concepts development, state of the art, and new directions. *International Journal of Cognitive Therapy*, 7(2):192–210.

David Feil-Seifer and Maja J. Mataric. 2005. Defining socially assistive robotics. In *9th International Conference on Rehabilitation Robotics, 2005. ICORR 2005.*, pages 465–468. IEEE.

Claire F Garandeau, Ihno A Lee, and Christina Salmivalli. 2014. Differential effects of the kiva anti-bullying program on popular and unpopular bullies. *Journal of Applied Developmental Psychology*, 35(1):44–50.

Edgar Omar Cebolledo Gutierrez and Olga De Troyer. 2014. Simbully: A'bullying in schools' simulation. In *FDG*.

Lynn Hawkins, Debra Pepler, and Wendy Craig. 2001. Naturalistic observations of peer interventions in bullying. *Social development*, 10(4):512–527.

Sheryl A. Hemphill, Aneta Kotevski, Michelle Tollit, Rachel Smith, Todd I. Herrenkohl, John W. Toumbourou, and Richard F. Catalano. 2012. Longitudinal predictors of cyber and traditional bullying perpetration in australian secondary school students. *Journal of Adolescent Health*, 51(1):59–65.

Sameer Hinduja and Justin W. Patchin. 2008. Cyberbullying: An exploratory analysis of factors related to offending and victimization. *Deviant behavior*, 29(2):129–156.

Adam O Horvath, AC Del Re, Christoph Flückiger, and Dianne Symonds. 2011. Alliance in individual psychotherapy. *Psychotherapy*, 48(1):9.

Alexander V. Libin and Elena V. Libin. 2004. Person-robot interactions from the robopsychologists' point of view: The robotic psychology and robotherapy approach. *Proceedings of the IEEE*, 92(11):1789–1803.

Daniel J Martin, John P Garske, and M Katherine Davis. 2000. Relation of the therapeutic alliance with outcome and other variables: a meta-analytic review. *Journal of consulting and clinical psychology*, 68(3):438.

Kathryn Modecki, Jeannie Minchin, Allen G. Harbaugh, Nancy G. Guerra, and Kevin C. Runions. 2014. Bullying prevalence across contexts: A meta-analysis measuring cyber and traditional bullying. *Journal of Adolescent Health*, 55(5):602–611.

Tonja R. Nansel, Mary D. Overpeck, Denise L. Haynie, W. June Ruan, and Peter C. Scheidt. 2003. Relationships between bullying and violence among us youth. *Archives of Pediatrics & Adolescent Medicine*, 157(4):348–353.

Dan Olweus. 1993. Bully/victim problems among schoolchildren: Long-term consequences and an effective intervention program. *Mental disorder and crime*, pages 317–349.

Dan Olweus and Susan P. Limber. 2010. Bullying in school: evaluation and dissemination of the olweus bullying prevention program. *American Journal of Orthopsychiatry*, 80(1):124.

Kirrily Pells, M.J. Ogando Portela, Patricia Espinoza Revollo, et al. 2016. Experiences of peer bullying among adolescents and associated effects on young adult outcomes: Longitudinal evidence from ethiopia, india, peru and viet nam. *UNICEF Office of Research-Innocenti, Florence*.

Rashmi Shetgiri. 2013. Bullying and victimization among children. *Advances in pediatrics*, 60(1):33.

Mona E. Solberg and Dan Olweus. 2003. Prevalence estimation of school bullying with the olweus bully/victim questionnaire. *Aggressive Behavior: Official Journal of the International Society for Research on Aggression*, 29(3):239–268.

Kim Storey, Ron Slaby, Melanie Adler, Jennifer Minotti, and Rachel Katz. 2013. Eyes on bullying toolkit. what can you do?

Mark Ter Maat, Khiet P. Truong, and Dirk Heylen. 2010. How turn-taking strategies influence users impressions of an agent. In *International Conference on Intelligent Virtual Agents*, pages 441–453. Springer.

Maria M. Ttofi and David P. Farrington. 2011. Effectiveness of school-based programs to reduce bullying: A systematic and meta-analytic review. *Journal of Experimental Criminology*, 7(1):27–56.

UNESCO. 2017. School violence and bullying: Global status report.

Cynthia Van Hee, Gilles Jacobs, Chris Emmery, Bart Desmet, Els Lefever, Ben Verhoeven, Guy De Pauw, Walter Daelemans, and Véronique Hoste. 2018. Automatic detection of cyberbullying in social media text. *PloS one*, 13(10):e0203794.

Jing Wang, Ronald J. Iannotti, and Jeremy W. Luk. 2012. Patterns of adolescent bullying behaviors: Physical, verbal, exclusion, rumor, and cyber. *Journal of school psychology*, 50(4):521–534.

Scott Watson, Natalie Vannini, Megan Davis, Sarah Woods, Marc Hall, Lynne Hall, and Kerstin Dautenhahn. 2007. Fearnot! an anti-bullying intervention: Evaluation of an interactive virtual learning environment. *Artificial Intelligence and Simulation of Behaviour (AISB), April*, 24.