# MOSafely: Building an Open-Source HCAI Community to Make the Internet a Safer Place for Youth

Xavier V. Caddle
University of Central Florida
Orlando, U.S.A
xavier.caddle@knights.ucf.edu

Afsaneh Razi
University of Central Florida
Orlando, Florida, U.S.A
afsaneh.razi@knights.ucf.edu

Seunghyun Kim
Georgia Institute of Technology
Atlanta, Georgia, U.S.A
seunghyun.kim@gatech.edu

Shiza Ali
Boston University
Boston, Massachusetts, U.S.A
shiza@bu.edu

Temi Popo
Mozilla Foundation
Canada
eyitemi@mozillafoundation.org

Gianluca Stringhini
Boston University
Boston, Massachusetts, U.S.A
gian@bu.edu

Munmun De Choudhury
Georgia Institute of Technology
Atlanta, Georgia, U.S.A
munmund@gatech.edu

Pamela J. Wisniewski
University of Central Florida
Orlando, Florida, U.S.A
pamwis@ucf.edu

## ABSTRACT

The goal of this one-day workshop is to build an active community of researchers, practitioners, and policy-makers who are jointly committed to leveraging human-centered artificial intelligence (HCAI) to make the internet a safer place for youth. This community will be founded on the principles of open innovation and human dignity to address some of the most salient safety issues of modern-day internet, including online harassment, sexual solicitation, and the mental health of vulnerable internet users, particularly adolescents and young adults. We will partner with Mozilla Research Foundation to launch a new open project named "MOSafely.org," which will serve as a platform for code library, research, and data contributions that support the mission of internet safety. During the workshop, we will discuss: 1) the types of contributions and technical standards needed to advance the state-of-the art in online risk detection, 2) the practical, legal, and ethical challenges that we will face, and 3) ways in which we can overcome these challenges through the use of HCAI to create a sustainable community. An end goal of creating the MOSafely community is to offer evidence-based, customizable, robust, and low-cost technologies that are accessible to the public for youth protection.

## CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; User studies; • **Social and professional topics** → **Computing / technology policy**.


*CSCW '21 Companion, October 23–27, 2021, Virtual Event, USA*

ACM ISBN 978-1-4503-8479-7/21/10.
https://doi.org/10.1145/3462204.3481731

## KEYWORDS

Adolescent Online Safety, HCAI, Open-Source Initiative, Risk Detection

**ACM Reference Format:**
Xavier V. Caddle, Afsaneh Razi, Seunghyun Kim, Shiza Ali, Temi Popo, Gianluca Stringhini, Munmun De Choudhury, and Pamela J. Wisniewski. 2021. MOSafely: Building an Open-Source HCAI Community to Make the Internet a Safer Place for Youth. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing (CSCW '21 Companion), October 23–27, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3462204.3481731

## 1 INTRODUCTION

This workshop seeks to leverage human-centered principles and innovative machine learning and artificial intelligence techniques to keep youth safe online. The general approach of using machine learning to detect online risk behaviors is not new. Yet, the bulk of the innovation in this space stays locked within academic research papers or behind corporate walls. We intend to unlock this potential. We will do this by bringing together a multidisciplinary and multi-organizational group of researchers, industry professionals, clinicians, and civil servants to research, build, evaluate, and bring to market state-of-the-art technologies that detect risk behaviors of youth online and/or their unsafe online interactions with others (e.g., cyberbullying, sexual solicitations and grooming, exposure to explicit content, non-suicidal self-injury, suicidal ideation, and other imminent risks). Our intention is to maximize societal impact by centralizing and making our open-source contributions widely available to the public to address youth online safety directly within the platforms that online risks are most likely to occur. As such, our open-source community building initiative, Modus Operandi Safely ("MOSafely"), will serve multiple end users that include, but are not limited to, social media platforms, youth safety coalitions, and other internet-based intermediaries (e.g., Apple iOS and Android smart devices, multi-player gaming platforms, internet service providers), who desire to proactively protect youth from serious online risks.

## 2 WORKSHOP GOALS

The primary goal of the workshop will be community building. This workshop will serve as the inaugural launch of MOSafely.org, an open-source community that leverages evidence-based research, data, and HCAI to help youth engage more safely online. The name "Modus Operandi Safely" (i.e., MOSafely) stems from our desire to help youth engage "more safely" online. As an open-source initiative, we have partnered with Mozilla to learn from their extensive experience creating open-source solutions. From this partnership we have learned the importance of have being supported by a diverse, committed team, and solidified our desire to work with a community to provide an online risk-detection platform. Towards this end, we will bring together a diverse group of researchers, industry professionals, youth service providers, and policy makers who have demonstrated a commitment to the mission of youth online safety and well-being, open innovation, and/or HCAI for youth risk detection in online contexts. Attendees will help us identify key stakeholders, best practices, challenges, and solutions for establishing the MOSafely community as an open source leader in the HCAI community for youth risk detection and online safety by addressing the following workshop themes.

## 3 WORKSHOP THEMES

**Theme 1**: *Approaches for Improving Online Risk Detection for Youth.* Following the rapid growth of social media, youth are increasingly exposed to harmful content and interactions online, ranging from pornography to offensive messages through online communities [8, 22, 26]. We would like to call the community to discuss the technical standards needed to advance the state-of-the-art in online risk detection. This would include but not limited to various techniques that could be implemented to further improve the existing online risk detection systems specifically geared towards youth as well as way to stimulate participation within the community. We raise the following questions:

(1) *How can we devise more sophisticated detection approaches that would detect multi-modal online risks through textual, visual, and meta data?*
(2) *What technical standards are needed for the centralized development of online risk detection algorithms for youth?*
(3) *What types of contributions (e.g., code libraries, evidence-based research, data sets, etc.) are needed to advance the state-of-the-art in the algorithmic risk detection of youth online risks?*

**Theme 2**: *Practical, Ethical, and Legal Considerations When Creating and Deploying Algorithms for Youth Risk Detection.* Developing machine learning models for online risk detection entails practical, legal, and ethical challenges that need to be taken into account. Ensuring the protection of the vulnerable populations we are trying to serve is mission critical to our approach. Often, algorithmic research can fall short if not considering the ethical implications of scraping, analyzing, and making classifications based on users' social media data [17]. When using such data specifically related to youth who are minors, there are numerous aspects that need to be considered such as consent, assent, and reporting incidents of child abuse and/or pornography [3]. In the past decade, social media has amassed a lot of data from youth, but the accessibility of said data has been limited; recently, there have been movement towards making it available such as shown in [5, 21]. In this theme we want to explore the practical, legal, and ethical challenges we will face using AI in online risk detection for the explicit purpose of protecting youth.

(1) *What are the legal and ethical implications of collecting the digital trace data of youth?*
(2) *How can the community be mobilized to work on detecting risks targeted towards youth online without exposing their data to the entire community? What infrastructure must be in place to safely collect and use teen data for risk detection?*
(3) *How can bias be avoided in youth online risk detection algorithms?*
(4) *What are the potential unintended consequences of developing and making widely available algorithms that detect youth risk behavior online?*

**Theme 3**: *Why Human-Centerdness is Needed in AI.* It is easy for Computer Scientists to focus on functionality and performance while developing computer algorithms. However, the HCI research community has identified the focus on such metrics without the incorporation of the human context to be unwise. As such, the need to have a human-centered approach to algorithm design has been highlighted in recent literature [2, 4, 12]. It is important to integrate human-centeredness in the development of MOSafely solutions to ensure transparency, explainability, and accountability. In this theme, we want to explore the need to have a Human-centered lens during risk detection algorithm development for youth online risk detection. These contributions include but are not limited to utilizing HCML and HCI methods in different cycles of developing AI systems for risk detection and online safety, dataset creation and design, developing and evaluating the systems, how to create ethical systems that take the most care of youth's privacy, and how to remove various types of bias from systems, and technical ML contributions. Thus, we pose the following questions:

(1) *How do we incorporate different stakeholders' perspectives and needs in the outputs of MOSafely?*
(2) *How do the current algorithm design techniques fall short in being user and stakeholder centered?*
(3) *What would be the key aspects of human-centeredness in machine learning that we should consider when trying to overcome these limitations?*
(4) *How can the incorporation of a human-centered viewpoint during algorithm design and development become a focal point in our community moving forward?*

## 4 CALL FOR PARTICIPATION

We will host a one-day virtual workshop with 40 to 60 participants from academia, industry, and civil society. To ensure a balanced mix of participants from HCI, design, social sciences, and other interdisciplinary fields, we will recruit participants via social media, social media groups (e.g., CHIMeta, CSCWMeta, CRA-WP), email list-servs, and appropriate community boards. We will also actively recruit participants from industry and civil society who are concerned about online safety issues and are interested contributing to the mission of MOSafely. This broad range of stakeholders will allow us to understand the needs and goals of our potential community members (i.e., contributors) as well as coalesce a large pool of

potential collaborators with which to engage during the workshop and beyond.

Workshop papers will be accepted through the MOSafely.org website: https://www.mosafely.org/workshops/cscw2021. At least one author of each accepted position paper must attend the workshop. Per SIGCHI conference guidelines, all participants must register for both the workshop and for at least one day of the conference. Workshop participants are asked to submit a brief statement of interest to ensure that their workshop participation is well-aligned with the workshop goals. Submissions can be structured in multiple ways: (1) Short bio's of each attendee with a statement of motivation/interest for attending the workshop, (2) an academic position paper (2-4 pages) in the SIGCHI extended abstract format discussing one or more of the workshop themes, or (3) a case study on relevant work that demonstrates a contribution towards HCAI/AI for youth online safety/risk detection. We also encourage potential attendees to explicitly state their commitment in joining MOSafely as a meaningful contributor that can help build and sustain the open-source community. We encourage submissions that are honest and subversive. Note that participants need not have prior experience with this type of work. Each submissions will be peer-reviewed by two program committee members and accepted based on the quality of the submission, relevance of the topic, and the diversity of the individual(s)' and their ability to meaningfully contribute to the workshop discussions and goals.

## 5 EXPECTED WORKSHOP CONTRIBUTIONS AND BEYOND

The expected outcome of the workshop will be a co-created agenda for establishing an inaugural community of MOSafely contributors who will play an active role in creating community standards, contributing code libraries and research, as well as taking on other leadership positions that support the community's mission. After the workshop, the organizers will invite workshop attendees to join the MOSafely open-source community and will report the workshop outcomes in a blog post on the MOSafely.org website. In terms of long-term outcomes, the MOSafely community will support two inter-related initiatives:

- An open source project that releases untrained algorithms relevant to youth online risk detection to the public as a way to gain market visibility and broad participation, so that others can train the algorithms with their own data sets and contribute code and expertise to as part of this open source project.
- A commercial Software as a Service (SaaS) Application Protocol Interface (API) that combines these algorithms into an easy-to-use and accessible service for online risk detection and mitigation.

The open source platform will provide typical community building resources, including contribution guidelines, issue tracking, documentation, and development resources. The project will initially be maintained by the workshop organizers with additional contributors gaining administrative roles as they contribute to the mission of MOSafely. Results from research generated by the community and code contributions from the open source project will be used to continuously improve the SaaS API. Developers may build product solutions by integrating open source code libraries, and online platforms could leverage the MOSafely SaaS API to detect and mitigate online risks that are facilitated through their platforms. Our intention is that this approach will broaden participation and create a shared societal responsibility of keeping youth safe online.

## 6 WORKSHOP CO-ORGANIZERS

**Xavier Caddle** is a PhD student at the University of Central Florida (UCF) and a member of the Socio-Technical Interaction Research (STIR) Lab. His current research focuses on conducting customer discovery and developing open-source standards and best practices for making MOSafely a sustainable community that leads the efforts for HCAI internet safety for youth.

**Afsaneh Razi** is a Ph.D. candidate at UCF and a member of the STIR Lab. Her recent works [13, 19] highlighted that online sexual experiences have become an irrevocable part of teens' sexual development. She discusses ethical challenges and considerations for data collection and development/deployment of adolescent online risk detection AI systems [1, 18, 20].

**Seunghyun Kim** is a Ph.D. student in the School of Interactive Computing at the Georgia Institute of Technology and a member of the Social Dynamics and Wellbeing Lab. His recent work highlighted the difference between the perspectives of the stakeholders of cyberbullying and its influence on cyberbullying detection algorithms [14].

**Shiza Ali** is a Ph.D. student at Boston University in the ECE Department. She is a member of the Security Lab (SeclaBU). Her research involves analyzing large datasets to understand malicious users online and developing mitigation techniques.

**Temi Popo** is an open innovation practitioner and creative technologist leading Mozilla's developer-focused strategy around Trustworthy AI and MozFest. She has worked across several industries in the area of Innovation and Strategic Foresight.

**Gianluca Stringhini** is an Assistant Professor in the ECE Department at Boston University and the Director of the SeclaBU Lab. He works in the area of data-driven security, applying computational techniques to make online users safe. For example, he has recently worked on mitigating coordinated online harassment [15, 16], cyberbullying [7], and disinformation [23, 27].

**Munmun De Choudhury** is an Associate Professor of Interactive Computing at Georgia Tech and the Director of the Social Dynamics and Well-Being Lab. She is best known for her work in laying the foundation of computational and human-centered techniques to responsibly and ethically employ social media in understanding and improving mental health [6, 9–11].

**Pamela Wisniewski** is an Associate Professor in the Department of Computer Science at the University of Central Florida and Director of the STIR Lab. Her research expertise lies at the intersection of social media, privacy, and online safety for adolescents (ages 13-17). She was one of the first researchers to recognize the need for resilience-based and teen-centric approaches for online safety and to back this stance up with empirical data [3, 19, 24–26].

## 7 PROGRAM COMMITTEE MEMBERS

The following individuals have confirmed their commitment to serving of the Program Committee. The responsibilities will include

reviewing 2-5 position papers/bios with statements of interest of potential workshop attendees, promoting the workshop within their personal networks, and if possible, attend the workshop to meaningfully contribute to the MOSafely mission:

- Zahra Ashktorab, Research Staff Member, IBM Research
- Jeremy Blackburn, Assistant Professor, Binghamton University
- Lindsay Blackwell, Senior Researcher, Twitter
- Laura Brown, Senior UX Researcher, Facebook
- Rosta Farzan, Associate Professor, University of Pittsburgh
- Ana Freire, Researcher and Lecturer, Pompeu Fabra University
- Shion Guha, Assistant Professor, University of Toronto
- Shirin Nilizadeh, University of Texas at Arlington
- Vivek Singh, Associate Professor, Rutgers University
- Kathryn Seigfried-Spellar, Associate Professor, Purdue University
- Thamar Solorio, Associate Professor, University of Houston
- Jacqueline Vickery, Associate Professor, University of North Texas

## ACKNOWLEDGMENTS

This research is supported by the U.S. National Science Foundation under grant #IIP-1827700. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. National Science Foundation.

## REFERENCES

[1] Zainab Agha, Neeraj Chatlani, Afsaneh Razi, and Pamela Wisniewski. 2020. Towards Conducting Responsible Research with Teens and Parents regarding Online Risks. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8.

[2] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.

[3] Karla Badillo-Urquiola, Zachary Shea, Zainab Agha, Irina Lediaeva, and Pamela Wisniewski. 2021. Conducting Risky Research with Teens: Co-Designing for the Ethical Treatment and Protection of Adolescents. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW3, Article 231 (Jan. 2021), 46 pages. https://doi.org/10.1145/3432930

[4] Eric PS Baumer. 2017. Toward human-centered algorithm design. *Big Data & Society* 4, 2 (2017), 2053951717718854. https://doi.org/10.1177/2053951717718854 arXiv:https://doi.org/10.1177/2053951717718854

[5] Ian Cairns. 2020. *Introducing a new and improved Twitter API.* Retrieved February 10, 2021 from https://blog.twitter.com/developer/en_us/topics/tools/2020/introducing_new_twitter_api.html

[6] Stevie Chancellor, Michael L Birnbaum, Eric D Caine, Vincent MB Silenzio, and Munmun De Choudhury. 2019. A taxonomy of ethical tensions in inferring mental health states from social media. In *Proceedings of the conference on fairness, accountability, and transparency*. 79–88.

[7] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the 2017 ACM on web science conference*. 13–22.

[8] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*. IEEE, 71–80.

[9] Munmun De Choudhury, Scott Counts, Eric J Horvitz, and Aaron Hoff. 2014. Characterizing and predicting postpartum depression from shared facebook data. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 626–638.

[10] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 7.

[11] Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 2098–2110.

[12] Rebecca Fiebrink and Marco Gillies. 2018. Introduction to the Special Issue on Human-Centered Machine Learning. *ACM Trans. Interact. Intell. Syst.* 8, 2 (June 2018), 7:1–7:7. https://doi.org/10.1145/3205942

[13] Heidi Hartikainen, Afsaneh Razi, and Pamela Wisniewski. 2021. Safe Sexting: The Advice and Support Adolescents Receive from Peers regarding Online Sexual Risks. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–31.

[14] Seunghyun Kim, Afsaneh Razi, Gianluca Stringhini, Pamela J Wisniewski, and Munmun De Choudhury. 2021. You Don't Know How I Feel: Insider-Outsider Perspective Gaps in Cyberbullying Risk Detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 15. 290–302.

[15] Chen Ling, Utkucan Balcı, Jeremy Blackburn, and Gianluca Stringhini. 2021. A first look at zoombombing. In *IEEE Symposium on Security and Privacy*.

[16] Enrico Mariconti, Guillermo Suarez-Tangil, Jeremy Blackburn, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Jordi Luque Serrano, and Gianluca Stringhini. 2019. " You Know What to Do" Proactive Detection of YouTube Videos Targeted by Coordinated Hate Attacks. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–21.

[17] Desmond U Patton, William R Frey, Kyle A McGregor, Fei-Tzin Lee, Kathleen McKeown, and Emanuel Moss. 2020. Contextual Analysis of Social Media: The Promise and Challenge of Eliciting Context in Social Media Posts with Natural Language Processing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 337–342.

[18] Afsaneh Razi, Zainab Agha, Neeraj Chatlani, and Pamela Wisniewski. 2020. Privacy Challenges for Adolescents as a Vulnerable Population. In *Networked Privacy Workshop of the 2020 CHI Conference on Human Factors in Computing Systems*.

[19] Afsaneh Razi, Karla Badillo-Urquiola, and Pamela J Wisniewski. 2020. Let's Talk about Sext: How Adolescents Seek Support and Advice about Their Online Sexual Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[20] Afsaneh Razi, Seunghyun Kim, Munmun De Choudhury, and Pamela Wisniewski. 2019. Ethical considerations for adolescent online risk detection AI systems. In *Good Systems: Ethical AI for CSCW (The 22nd ACM Conference on Computer-Supported Cooperative Work and Social Computing)*.

[21] Naomi Shiffman. 2021. *Social Media, Social Life. Teens Reveal Their Experiences.* Retrieved February 18, 2021 from https://help.crowdtangle.com/en/articles/4558716-understanding-and-citing-crowdtangle-data

[22] Muhammad Uzair Tariq, Afsaneh Razi, Karla Badillo-Urquiola, and Pamela Wisniewski. 2019. A Review of the Gaps and Opportunities of Nudity and Skin Detection Algorithmic Research for the Purpose of Combating Adolescent Sexting Behaviors. In *Human-Computer Interaction. Design Practice in Contemporary Societies (Lecture Notes in Computer Science)*, Masaaki Kurosu (Ed.). Springer International Publishing, Cham, 90–108. https://doi.org/10.1007/978-3-030-22636-7_6

[23] Yuping Wang, Fatemeh Tamahsbi, Jeremy Blackburn, Barry Bradlyn, Emiliano De Cristofaro, David Magerman, Savvas Zannettou, and Gianluca Stringhini. 2021. Understanding the Use of Fauxtography on Social Media. In *AAAI International Conference on Web and Social Media (ICWSM)*.

[24] Pamela Wisniewski, Arup Kumar Ghosh, Heng Xu, Mary Beth Rosson, and John M. Carroll. 2017. Parental Control vs. Teen Self-Regulation: Is There a Middle Ground for Mobile Online Safety?. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 51–69. https://doi.org/10.1145/2998181.2998352

[25] Pamela Wisniewski, Haiyan Jia, Na Wang, Saijing Zheng, Heng Xu, Mary Beth Rosson, and John M. Carroll. 2015. Resilience Mitigates the Negative Effects of Adolescent Internet Addiction and Online Risk Exposure. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 4029–4038. https://doi.org/10.1145/2702123.2702240

[26] Pamela Wisniewski, Heng Xu, Mary Beth Rosson, Daniel F. Perkins, and John M. Carroll. 2016. Dear Diary: Teens Reflect on Their Weekly Online Risk Experiences. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 3919–3930. https://doi.org/10.1145/2858036.2858317

[27] Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2019. Disinformation warfare: Understanding state-sponsored trolls on Twitter and their influence on the web. In *Companion proceedings of the 2019 world wide web conference*. 218–226.