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

XAVIER CADDLE, University of Central Florida, U.S.A
AFSANEH RAZI, University of Central Florida, U.S.A
SEUNGHYUN KIM, Georgia Institute of Technology, U.S.A
SHIZA ALI, Boston University, U.S.A
TEMI POPO, Mozilla Foundation, Canada
GIANLUCA STRINGHINI, Boston University, U.S.A
MUNMUN DE CHOUDHURY, Georgia Institute of Technology, U.S.A
PAMELA WISNIEWSKI, University of Central Florida, U.S.A

The goal of this one-day workshop is to begin building 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 solutions that are accessible to the public for the purpose of youth protection.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**; User studies.

Additional Key Words and Phrases: Adolescent Online Safety, HCAI, Open-Source Initiative, Risk Detection

ACM Reference Format:
Xavier Caddle, Afsaneh Razi, Seunghyun Kim, Shiza Ali, Temi Popo, Gianluca Stringhini, Munmun De Choudhury, and Pamela Wisniewski. 2021. MOSafely: Building an Open-Source HCAI Community to Make the Internet a Safer Place for Youth. In *Proceedings of The 24TH ACM Conference On Computer-Supported Cooperative Work And Social Computing (CSCW '24).* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

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

CSCW '24, October 23–27, 2021, Virtual

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

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.

During our one-day workshop attendees will work together to address the following high-level themes: 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 Human Centered Artifical Intelligence (HCAI) for online risk detection to create a sustainable community that has the potential to change the world. By addressing these themes together as a community, the goal of this workshop is to start actively building a vibrant ecosystem of contributors that shape and sustain MOSafely's mission of leveraging open innovation and HCAI for the purpose of protecting youth online.

2 BACKGROUND

2.1 The Importance of Promoting Online Safety for Youth and Young Adults

Depressive symptoms, non-suicidal self-injury, and suicide rates have increased significantly among youth (ages 13-17) and young adults (ages 18-23) in the last decade [34, 49] with recent research suggesting that the rise in social media use is a contributing factor to this negative trend [48]. Indeed, the majority of risks youth are exposed to online occur via social media sites [27], and the combination of social media and personal digital devices has created a problematic situation by providing unmediated internet access and a potentially dangerous level of "practical obscurity" [7], or limited visibility, into the risky activities teens engage in online. For the most part, the burden of protecting teens from online risks has traditionally fallen on the shoulders of parents [29, 36]; however, by launching MOSafely, we aim to create a significant societal pivot, where youth online safety becomes a shared responsibility for all, especially the online platforms in which teens encounter risks, thereby serving teens, their families, and society as a whole.

2.2 Leveraging Open Innovation and HCAI for the Online Protection of Youth

Most commercially available risk detection solutions focus on detecting objectionable or illegal content, as opposed to detecting online risks for the explicitly purpose of safeguarding youth. However, Facebook and social media platforms have been at the forefront of developing proprietary algorithms to detect risk behaviors, such as suicidal intentions [42] and cyberbullying [5]. While considered the state-of-the-art, these risk detection algorithms are platform specific, opaque as to how they work, not publicly available for reuse, and typically focus on a narrow range of the most serious risk behaviors, ignoring important patterns of risk escalation that could help circumvent severe harm before it occurs [25]. In addition, these approaches do not fully incorporate the perspectives of diverse stakeholders, especially those of victims, resulting in missing implicit references as well as atypical framings of the online risk [26]. Most of the current algorithms for risk detection also suffer from a high number of false positives [13, 40], making them fairly unusable in real world settings. False positive rates are high because these algorithms often do not leverage human-centered approaches that leverage insights gained from contextual information (e.g., metadata, such as the time of day a message was sent, or whom sent or received the message), patterns of behavior over time (e.g., sexual grooming patterns [6]).

These two limitations (i.e., opaqueness and limited availability of proprietary algorithms combined with the lack of human-centeredness in the development process) combined create a compelling case for leveraging open innovation and HCAI to build better solutions for the protection of youth online. The inner workings of proprietary algorithms are hidden from the general public and thus restrict us from understanding how they work and at times stop us from using or creating solutions which utilize their functionality [42]. Open-innovation has at its core the concept of using both internal and external ideas to create solutions [12], thus we believe that it can aid in solving this issue of proprietary blocks as we desire to work with the community to create open-source algorithmic solutions.Further, the integration of HCAI into this problem space is critical. HCAI looks at the AI and ML algorithms through the lens of humans, advocating that such systems need incorporate socio-cultural understanding of humans as well as help humans understand AI [41]. Therefore, we present the following goals and themes for our workshop.

3 WORKSHOP GOALS

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.

A strong community-based commitment is key to the success of MOSafely. Thus, the primary goal of the workshop will be community building. 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. We will build upon previous CSCW and CHI workshops [17, 20, 22, 24, 28, 33, 45, 54], which addressed related themes in keeping with our efforts. 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.

4 WORKSHOP THEMES

The themes of our workshop were specifically chosen to bring about discussion on community building to support online risk detection for youth.

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 [11, 47, 53]. Past literature on online risk detection algorithms has adapted approaches from machine learning and natural language processing. Most approaches for sexual risk detection are currently based on traditional ML algorithms while recently researchers utilized deep learning models [18]. Cyberbullying detection studies have also implemented supervised learning techniques [43]; however, obtaining large amounts of data as well as the transparency of these models are recurring challenges of such approach [44]. Currently due to challenges on collecting sensitive data from youth for the purpose of online risk detection, most researchers rely on unrealistic or general data for risk detection

[47]. Therefore, establishing ecologically valid training datasets of teen's digital trace data and defining and quantifying risks for having informed ground truth for machine learning systems are important.

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 [35]. 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 [8, 46]. 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, 19]. There have been issues particularly related to bias, stereotyping, and marginalization in systems using these technologies [32]. Thus, it is important to integrate human-centeredness in the development of MOSafely solutions to ensure transparency, explainability, and accountability. Embedding human-centeredness in the development of the online youth safety system will help ensure the robustness of the open-source tool and also enlarge the potential usages of MOSafely in protecting the safety of youth online.

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. For this theme, 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?*

5 PARTICIPANTS

We propose a one-day virtual workshop with 40 to 60 participants from academia, industry, and civil society. Participants will be recruited from the broader ACM and ACM SIGCHI communities, the extended research networks of the workshop organizers, and HCAI and AI researchers and practitioners across multiple disciplines. 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. These efforts will also be supported by the workshop's program committee. 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.

6 CALL FOR PARTICIPATION

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. We invite and encourage submissions from researchers from academia, industry, non-profits, and governments (national, regional, local, tribal), and welcome a wide range of disciplinary perspectives. We also strongly promote the inclusion of women, people of color, and other diverse voices that can offer their unique perspectives on protecting vulnerable populations from online risks, violence, and abuse. Each submissions will be peer-reviewed by two program committee members and will be 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.

Workshop papers will be accepted through a web-based submission form on 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.

7 WORKSHOP ACTIVITIES

The workshop will be held over one (1) day. The structure will lend itself towards discussion on how individual siloed risk detection efforts can come together as a strong community to create solutions which keep teens safe online. Participants are encouraged identify areas where teens are exposed to risky content online and discuss how we as a community can mitigate this issue. We plan to facilitate 20 participants.

(1) **Welcome/Introductions** (15 minutes): The organizers will introduce themselves and the mission behind creating the MOSafely open-source community. They will briefly cover logistics, including the workshop schedule and high-level goals of the workshop.
(2) **Lightning Talks** (1.5 hours): Attendees will introduce themselves and briefly present their position or work relevant to the workshop. Lightning talks should loosely align with Theme 1 of the workshop on the types of contributions necessary for advancing the state-of-the art in HCAI for promoting online safety and risk mitigation of youth.
(3) **Break**: 15 minute break.
(4) **Keynote Speaker** (1 hour): To inspire participants and spark discussion, we will have Temi Popo, Who leads developer-focused strategies around Trustworthy AI at Mozilla and is well-versed in the paradigm of Open Leadership. There will be an opportunity for workshop participants to engage with Ms. Popo in a Q&A session after her keynote.
(5) **Lunch** (1 hour): During lunch, we may use a social gathering platform such as GatherTown for workshop participants to engage with one another on a more personal level.
(6) **Large Group Discussion** (1 hour): Workshop participants will brainstorm and identify potential challenges that will need to be addressed for creating a sustainable and vibrant community of scholars, practitioners, civil servants, and policy-makers committed to leveraging evidence-based research and advances in HCAI to translate open innovation into real-world practice for protecting youth online. This discussion aligns with Theme 2 of the workshop.
(7) **Break**: 15 minute break.
(8) **Break-out Activities** (1.5 hours): Participants will breakout into smaller groups to create actionable solutions and tangible project-plans for tackling these challenges (Theme 3). Potential breakout groups may include: 1) Specific HCAI approaches participants have developed for specific online risk contexts (e.g., harassment, abuse, sexual solicitations, mental health risks, etc.), 2) Important HCAI concerns around trustworthiness, explainability, transparency, and accountability, and 2) Broader around ethical and legal implications of developing such technologies and making them accessible to the general public. Breakout groups will be formed based on emerging themes in submissions of accepted attendees and based on discussions raised during the first half of the workshop.
(9) **Reporting Outcomes** (45 minutes): Each small group will report back their ideas to the larger group. Participants may use several approaches to communicate their ideas, such developing project proposals, design fictions, mind-maps, or architecture diagrams. Each group will have an opportunity to get feedback from other workshop participants.
(10) **Next Steps** (30 minutes): The workshop will conclude with the organizers synthesizing the discussions and outcomes from the workshop and brainstorming with attendees on necessary next steps for officially launching MOSafely after the conclusion of the workshop.

8 EXPECTED WORKSHOP CONTRIBUTIONS AND BEYOND

The expected outcome of the workshop will be a co-created agenda for officially 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. Based on the preference of attendees, we will also create a listserv or forum for community-based organizing and ask workshop attendees to invite individuals from their extended networks to grow the MOSafely community. 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. Ultimately, the MOSafely community will provide these resources to developers and small to mid-sized internet-based companies that cater to youth. 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.

9 WORKSHOP WEBSITE

The website for the workshop is https://www.mosafely.org/workshops/cscw2021. All information related to the workshop (e.g., call for participation, important dates, schedule) will be available on this website upon acceptance of the workshop. Once participants have been selected, the website will also host the accepted submissions with the permission of the authors.

10 EQUIPMENT AND SUPPLIES

In keeping with the CSCW 2021 virtual meeting requirements, this workshop will be held using Zoom conferencing. Zoom breakout rooms will be utilized to facilitate smaller group discussions. We will also consider using virtual community engagement platforms, such as GatherTown [21], depending on the expressed interests of our attendees.

11 WORKSHOP CO-ORGANIZERS

The MOSafely workshop co-organizers are the PI/Co-PIs (and their students) on a National Science Foundation (NSF) Partnerships for Innovation (PFI) grant that funds this initiative.

Xavier Caddle is a PhD student in the Department of Computer Science 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 in the Department of Computer Science at UCF and a member of the STIR Lab. Her dissertation research is aimed at improving adolescent online safety by utilizing human-centered insights and machine learning to detect unwanted sexual risk experiences of adolescents. Her recent works [23, 38] highlighted that online sexual experiences have become an irrevocable part of teens' sexual development and identified the benefits and challenges when youth seek support and receive support for these experiences. In her work she discusses ethical challenges and considers for data collection and development/deployment of adolescent online risk detection AI systems [1, 37, 39].

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 research is focused on working to develop human-centered machine learning algorithms to assess online risk (e.g., cyberbullying, harassment, abuse, and self-harm). His recent work highlighted the difference between the perspectives of the stakeholders of cyberbullying and its influence on cyberbullying detection algorithms [26].

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. Her recent work involves developing tools to reduce cyberbullying and sexual harassment online, specifically when targeted towards teens.

Temi Popo is an open innovation practitioner and creative technologist leading Mozilla's developer-focused strategy around Trustworthy AI and MozFest. In 2012, she graduated magna cum laude from Mount Holyoke College, where she studied International Relations and Digital Media (computer science and film production). Ms. Popo also holds a Master's in Digital Experience Innovation from the University of Waterloo, with professional certification in digital publishing from NYU. 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 is a Senior Personnel for the grant supporting this effort. Dr. Stringhini 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 [30, 31], cyberbullying [10], and disinformation [50, 55].

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 the Co-Primary Investigator of the grant supporting this effort. Dr. De Choudhury 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 [9, 14–16].

Pamela Wisniewski is an Associate Professor in the Department of Computer Science at the University of Central Florida and Director of the STIR Lab. She is the Primary Investigator of the grant supporting this effort. Dr. Wisniewski's 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, rather than an abstinence-based approaches, and to back this stance up with empirical data [3, 38, 51–53].

12 PROGRAM COMMITTEE MEMBERS

The following individuals have confirmed their commitment to serving of the Program Committee should our workshop be accepted. 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] Macy Bayern. 2017. *How AI became Instagram’s weapon of choice in the war on cyberbullying.* Retrieved June 7, 2021 from https://mashable.com/2017/11/28/facebook-ai-suicide-prevention-tools/

[6] Pamela J. Black, Melissa Wollis, Michael Woodworth, and Jeffrey T. Hancock. 2015. A linguistic analysis of grooming strategies of online child sex offenders: Implications for our understanding of predatory sexual behavior in an increasingly computer-mediated world. *Child Abuse Neglect* 44 (2015), 140–149. https://doi.org/10.1016/j.chiabu.2014.12.004

[7] Lindsay Blackwell, Emma Gardiner, and Sarita Schoenebeck. 2016. Managing Expectations: Technology Tensions among Parents and Teens. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work amp; Social Computing* (San Francisco, California, USA) *(CSCW ’16)*. Association for Computing Machinery, New York, NY, USA, 1390–1401. https://doi.org/10.1145/2818048.2819928

[8] 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

[9] 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.

[10] 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.

[11] 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.

[12] H.W. Chesbrough, Harvard Business School Press, and J.S. Brown. 2003. *Open Innovation: The New Imperative for Creating and Profiting from Technology.* Harvard Business School Press. https://books.google.com/books?id=4hTRWStFhVgC

[13] Mateus de Castro Polastro and Pedro Monteiro da Silva Eleuterio. 2010. NuDetective: A Forensic Tool to Help Combat Child Pornography through Automatic Nudity Detection. In *2010 Workshops on Database and Expert Systems*

Applications. 349–353. https://doi.org/10.1109/DEXA.2010.74

[14] 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.

[15] 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.

[16] 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.

[17] Michael A. Devito, Ashley Marie Walker, Jeremy Birnholtz, Kathryn Ringland, Kathryn Macapagal, Ashley Kraus, Sean Munson, Calvin Liang, and Herman Saksono. 2019. Social Technologies for Digital Wellbeing Among Marginalized Communities. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing* (Austin, TX, USA) *(CSCW '19)*. Association for Computing Machinery, New York, NY, USA, 449–454. https://doi.org/10.1145/3311957.3359442

[18] Mohammadreza Ebrahimi, Ching Y Suen, and Olga Ormandjieva. 2016. Detecting predatory conversations in social media by deep convolutional neural networks. *Digital Investigation* 18 (2016), 33–49.

[19] 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

[20] Kenneth R. Fleischmann, Sherri R. Greenberg, Danna Gurari, Abigale Stangl, Nitin Verma, Jaxsen R. Day, Rachel N. Simons, and Tom Yeh. 2019. Good Systems: Ethical AI for CSCW. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing* (Austin, TX, USA) *(CSCW '19)*. Association for Computing Machinery, New York, NY, USA, 461–467. https://doi.org/10.1145/3311957.3359437

[21] GatherTown. 2020. *Gather*. Retrieved June 7, 2021 from https://gather.town/

[22] Cole Gleason, Patrick Carrington, Lydia B. Chilton, Benjamin M. Gorman, Hernisa Kacorri, Andrés Monroy-Hernández, Meredith Ringel Morris, Garreth W. Tigwell, and Shaomei Wu. 2019. Addressing the Accessibility of Social Media. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing* (Austin, TX, USA) *(CSCW '19)*. Association for Computing Machinery, New York, NY, USA, 474–479. https://doi.org/10.1145/3311957.3359439

[23] 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.

[24] Kori Inkpen, Stevie Chancellor, Munmun De Choudhury, Michael Veale, and Eric P. S. Baumer. 2019. Where is the Human? Bridging the Gap Between AI and HCI. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3290607.3299002

[25] Haiyan Jia, Pamela J. Wisniewski, Heng Xu, Mary Beth Rosson, and John M. Carroll. 2015. Risk-Taking as a Learning Process for Shaping Teen's Online Information Privacy Behaviors. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work amp; Social Computing* (Vancouver, BC, Canada) *(CSCW '15)*. Association for Computing Machinery, New York, NY, USA, 583–599. https://doi.org/10.1145/2675133.2675287

[26] 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.

[27] Mitchell Kimberly, Jones Lisa, David Finkelhor, and Janis Wolak. 2014. *Teens, Technology and Romantic Relationships*. Retrieved May 31, 2017 from http://www.pewinternet.org/2015/10/01/teens-technology-and-romantic-relationships

[28] Min Kyung Lee, Nina Grgić-Hlača, Michael Carl Tschantz, Reuben Binns, Adrian Weller, Michelle Carney, and Kori Inkpen. 2020. Human-Centered Approaches to Fair and Responsible AI. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3334480.3375158

[29] Sook-Jung Lee. 2013. Parental restrictive mediation of children's internet use: Effective for what and for whom? *New Media & Society* 15, 4 (2013), 466–481. https://doi.org/10.1177/1461444812452412 arXiv:https://doi.org/10.1177/1461444812452412

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

[31] 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.

[32] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635* (2019).

[33] Michael Muller, Cecilia Aragon, Shion Guha, Marina Kogan, Gina Neff, Cathrine Seidelin, Katie Shilton, and Anissa Tanweer. 2020. Interrogating Data Science. In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing* (Virtual Event, USA) *(CSCW '20 Companion)*. Association for Computing Machinery, New York, NY, USA, 467–473. https://doi.org/10.1145/3406865.3418584

[34] National Institute of Mental Health. 2015. *Suicide*. Retrieved June 7, 2021 from https://www.nimh.nih.gov/health/statistics/suicide

[35] 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.

[36] Sonia Livingstone Ph.D. and Ellen J. Helsper Ph.D. 2008. Parental Mediation of Children's Internet Use. *Journal of Broadcasting & Electronic Media* 52, 4 (2008), 581–599. https://doi.org/10.1080/08838150802437396 arXiv:https://doi.org/10.1080/08838150802437396

[37] 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*.

[38] 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.

[39] 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)*.

[40] Kelly Reynolds, April Kontostathis, and Lynne Edwards. 2011. Using Machine Learning to Detect Cyberbullying. In *2011 10th International Conference on Machine Learning and Applications and Workshops*, Vol. 2. 241–244. https://doi.org/10.1109/ICMLA.2011.152

[41] Mark O Riedl. 2019. Human-centered artificial intelligence and machine learning. *Human Behavior and Emerging Technologies* 1, 1 (2019), 33–36.

[42] Rebecca Ruiz. 2017. *Facebook's AI suicide prevention tool can save lives, but the company won't say how it works.* Retrieved June 7, 2021 from https://mashable.com/2017/11/28/facebook-ai-suicide-prevention-tools/

[43] Semiu Salawu, Yulan He, and Joanna Lumsden. 2017. Approaches to automated detection of cyberbullying: A survey. *IEEE Transactions on Affective Computing* (2017).

[44] Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. 2019. *Explainable AI: interpreting, explaining and visualizing deep learning*. Vol. 11700. Springer Nature.

[45] Devansh Saxena, Erhardt Graeff, Shion Guha, EunJeong Cheon, Pedro Reynolds-Cuéllar, Dawn Walker, Christoph Becker, and Kenneth R. Fleischmann. 2020. Collective Organizing and Social Responsibility at CSCW. In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing* (Virtual Event, USA) *(CSCW '20 Companion)*. Association for Computing Machinery, New York, NY, USA, 503–509. https://doi.org/10.1145/3406865.3418593

[46] 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

[47] 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

[48] Jean M. Twenge, Thomas E. Joiner, Megan L. Rogers, and Gabrielle N. Martin. 2018. Increases in Depressive Symptoms, Suicide-Related Outcomes, and Suicide Rates Among U.S. Adolescents After 2010 and Links to Increased New Media Screen Time. *Clinical Psychological Science* 6, 1 (2018), 3–17. https://doi.org/10.1177/2167702617723376 arXiv:https://doi.org/10.1177/2167702617723376

[49] Alicia VanOrman and Beth Jarosz. 2016. *Suicide Replaces Homicide as Second-Leading Cause of Death Among U.S. Teenagers.* Retrieved June 7, 2021 from https://www.prb.org/resources/suicide-replaces-homicide-as-second-leading-cause-of-death-among-u-s-teenagers

[50] 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)*.

[51] 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

[52] 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

[53] 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

[54] Marisol Wong-Villacres, Aakash Gautam, Wendy Roldan, Lucy Pei, Jessa Dickinson, Azra Ismail, Betsy DiSalvo, Neha Kumar, Tammy Clegg, Sheena Erete, Emily Roden, Nithya Sambasivan, and Jason Yip. 2020. From Needs to Strengths: Operationalizing an Assets-Based Design of Technology. In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing* (Virtual Event, USA) *(CSCW '20 Companion)*. Association for Computing Machinery, New York, NY, USA, 527–535. https://doi.org/10.1145/3406865.3418594

[55] 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.