SHAI 2023: Workshop on Designing for Safety in Human-AI Interactions

Nitesh Goyal
teshgoyal@acm.org
Google Research, Google
New York, NY, United States

Sungsoo Ray Hong
shong31@gmu.edu
George Mason University
Fairfax, VA, United States

Regan L. Mandryk
regan@acm.org
University of Saskatchewan
Saskatoon, SK, Canada

Toby Jia-Jun Li
toby.j.li@nd.edu
University of Notre Dame
Notre Dame, IN, United States

Kurt Luther
kluther@vt.edu
Virginia Tech
Arlington, VA, United States

Dakuo Wang
dakuo@acm.org
Northeastern University
Boston, MA, United States

ABSTRACT

Generative ML models present a novel opportunity for a wider group of societal members to engage with AI, imagine new use cases, and applications with an increasing ability to disseminate the outcomes of such endeavors to larger audiences. However, owing to the novelty and despite best intentions, inadvertent outcomes might accrue leading to harms, especially to marginalized groups in society. As this field of Human AI Interaction advances, academic/industry researchers, and industry practitioners have an opportunity to brainstorm how to best utilize this new technology. Our workshop is aimed at such practitioners and researchers at the intersection of AI and HCI who are interested in collaboratively identifying challenges, and solutions to create safer outcomes with Generative ML models.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**; • **Computing methodologies → Artificial intelligence**; **Machine learning approaches**.

KEYWORDS

Human-AI interaction, Safety, Harms, Generative models, Generative AI, Responsible AI, LLM, AI, ML

ACM Reference Format:

Nitesh Goyal, Sungsoo Ray Hong, Regan L. Mandryk, Toby Jia-Jun Li, Kurt Luther, and Dakuo Wang. 2023. SHAI 2023: Workshop on Designing for Safety in Human-AI Interactions. In *28th International Conference on Intelligent User Interfaces (IUI '23 Companion), March 27–31, 2023, Sydney, NSW, Australia.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3581754.3584169

1 MOTIVATION

Machine learning (ML) models continue to evolve and are being deployed across multiple applications [4], including those in direct engagement with society (e.g., adaptive chatbots [19, 25], content moderation outcomes (e.g., https://perspectiveapi.com, etc.). Such advances are uniquely powerful as they now point to the ability for computing to produce generative interactions that can adapt to human engagement and motivations. These can be potentially benign motivations like assisting in writing academic papers [22] to predicting recidivism using algorithms [21] to outright causing harm to online communities by generating hate speech [23]. In general, as the models are being deployed for handling social-level of subjective and sensitive tasks, it is becoming crucial to not just understand how to detect and revise the models such that end users can leverage the AI-driven benefit in a safer environment, but focus on providing a solution that aids multiple stakeholders. However, detecting and fixing safety-related issues in AI can be deeply challenging for academic and practice communities alike. For example, there is no consensus formalized in the definition of safety in several domains, such as hate speech concerns [20]. Further, detecting such safety concerns requires iterative real-time discovery of patterns of abuse [17], and generating intensive appropriately annotated datasets to train models about such abuse [10, 15]. This is a whack-a-mole situation that remains unsolved. While multiple researchers continue to resolve this situation related to definitions and taxonomy, this workshop focuses also on how we can move ahead already and identify ways to address known challenges.

Detecting and managing safety concerns in AI is a multi-faceted challenge [5]. Amongst many facets, this will require iterative socio-technical understanding of human behavior that motivates such outcomes [6]; theoretical models and frameworks that define such behaviors [24]; sociological observation of impact of such human and algorithmic behaviors [16]; computational advances in pursuing research beyond model accuracy by focusing on catastrophic consequences to humans [1, 14]; creative opportunities for design to manage the user experience and journeys of humans who are likely to be targeted at scale [10]; practical challenges of tracking human and algorithmic harmful / unsafe operations at scale [13]; balancing model accuracy and safety-related metrics which pose a technical dilemma for product-oriented practitioners and ethicists [8, 9, 11, 12, 26, 27], and balancing safety with constructive conflict [3].

Therefore, the main research question of ***SHAI*** — Safety in Human-AI Interactions — is: ***How can we make the outcomes of ML models, especially generative models, safer when humans engage with these models?*** To achieve this, the workshop brings

IUI '23 Companion, March 27–31, 2023, Sydney, NSW, Australia

ACM ISBN 979-8-4007-0107-8/23/03.
https://doi.org/10.1145/3581754.3584169

multidisciplinary industry and academic partners who brainstorm and identify potential opportunities for collaboration to address one or more of the above-highlighted challenges. This would require researchers at the intersection of AI and HCI to learn from each other and arrive at a consensus on what gaps lie in our collective understanding of human and ML model functioning and focus on solutions. We hope that this workshop will be one of the many, focusing on safety and harms, enabling this community to interconnect and grow.

2 WORKSHOP ORGANIZERS

The workshop's organizers are diverse in their backgrounds and research expertise, combining both industry and academia. Their research expertise covers a wide range of subfields in HCI and AI, including Online Hate, Safety, Harms, AI Inequality, Bias in Machine Learning, Fairness and Inclusion, Crowdsourcing, CSCW and Social Computing, Human-AI Collaboration, Human-AI Interaction, Interactive Data annotation, Intelligent Design, Natural Language Processing, and, Visual Analytics. The short bios of every organizer are listed as follows:

Nitesh Goyal leads research on tools designed to build AI responsibly at Google Research. His work has focused on AI for social good for marginalized populations, including tools for journalists/activists to manage harassment, reducing biases during investigative sensemaking, unpacking the role of data annotators' identity on ML outcomes and more.

Sungsoo Ray Hong (he/him) is an Assistant Professor in the Department of Information Sciences and Technology at George Mason University. His research mission is Alignable AI, aiming at establishing empirical understanding and designing novel tools to make AI aligned to humans' expectations, norms, and mental model.

Regan L. Mandryk (she/her) is a Canada Research Chair in Digital Gaming Technologies and Experiences and Professor of Computer Science at the University of Saskatchewan. Her work focuses on how people use playful technologies for social and emotional wellbeing, and how toxicity thwarts the connection and recovery benefits provided by multiplayer games.

Toby Jia-Jun Li (he/him) is an Assistant Professor in Computer Science and Engineering at the University of Notre Dame. Toby designs, builds, and studies interactive systems that facilitate effective human-AI collaboration in various task domains. Several focus areas of his work include human-centered data science, human-AI co-creation in creative tools, human-AI collaboration in programming, and worker empowerment against AI inequality in gig work.

Kurt Luther (he/him) is an associate professor of computer science and (by courtesy) history at Virginia Tech. His research group, the Crowd Intelligence Lab, builds and studies systems that combine the complementary strengths of crowdsourced human intelligence and AI to support ethical, effective investigations.

Dakuo Wang (he/him) is a Senior Research Staff Member and leads the human-centered natural language interaction strategy at IBM Research. He specializes in designing and developing human-centered AI systems for real-world user needs and has published more than 50 papers and 50 patents on related topics.

3 WORKSHOP OVERVIEW

Workshop participants were IUI attendees, at the intersection of AI and HCI, whose research has relevance to developing not only accurate but also safe AI-driven applications and solutions. Participants submitted a Position Paper, an Opinion Paper, or a Late-breaking Work.

The Workshop was organized as a half day mini Conference in a hybrid format, allowing in-person and virtual participation. Workshop related materials, such as conference video, position papers, and discussion outcomes are provided on the workshop website. The workshop plan included an Ice breaking, Hands-on Exercise, and a Group Discussion. In total four papers were presented at the workshop.

Some of the authors discussed how Information theory can be leveraged from most natural applications of combinatorial creativity with modern generative AI for the safety-creativity tradeoff. While in [7], authors discussed that it is important to consider what is normal and abnormal when it comes to safety. This should include abnormal occurrences. Authors end their paper with an appeal to include normative considerations to govern what kind of variables and values are represented in a model yields highly intuitive results. Alternatively, in [18], authors focus on Explanation from the perspective of philosophy and now within AI research (Explainable AI/XAI) and point that this new development in XAI can benefit from incredibly long history and discourse in Philosophy domain. Finally, in [2], authors focus on a specific domain of career searching for members of community with Autism spectrum disorder (ASD). Authors aim to explore how future designers can leverage technology and AI to motivate ASDs to effectively collaborate in their career-seeking process.

REFERENCES

[1] Saleema Amershi, Max Chickering, Steven M Drucker, Bongshin Lee, Patrice Simard, and Jina Suh. 2015. Modeltracker: Redesigning performance analysis tools for machine learning. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 337–346.

[2] Zinat Ara and Sungsoo Ray Hong. 2023. Exploring Design Space of Collaborative Career-Seeking Experience for People on Autism Spectrum. (2023).

[3] Amanda Baughan, Ashwin Rajadesingan, Alexis Hiniker, Paul Resnick, and Amy Bruckman. 2022. SIG on Designing for Constructive Conflict. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, 1–2.

[4] Minsuk Choi, Cheonbok Park, Soyoung Yang, Yonggyu Kim, Jaegul Choo, and Sungsoo Ray Hong. 2019. Aila: Attentive interactive labeling assistant for document classification through attention-based deep neural networks. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.

[5] Alexandra Chouldechova and Aaron Roth. 2018. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810* (2018).

[6] Jerry Alan Fails and Dan R Olsen Jr. 2003. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*. 39–45.

[7] Laura Fearnley. 2023. Norms and Causation in Artificial Morality. (2023).

[8] Yuyang Gao, Tong Sun, Rishab Bhatt, Dazhou Yu, Sungsoo Hong, and Liang Zhao. 2021. Gnes: Learning to explain graph neural networks. In *2021 IEEE International Conference on Data Mining (ICDM)*. IEEE, 131–140.

[9] Yuyang Gao, Tong Steven Sun, Liang Zhao, and Sungsoo Ray Hong. 2022. Aligning eyes between humans and deep neural network through interactive attention alignment. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–28.

[10] Nitesh Goyal, Ian Kivlichan, Rachel Rosen, and Lucy Vasserman. 2022. Is Your Toxicity My Toxicity? Exploring the Impact of Rater Identity on Toxicity Annotation. In *Proceedings of ACM in Human Computer Interaction CSCW*.

[11] Anhong Guo, Ece Kamar, Jennifer Wortman Vaughan, Hanna Wallach, and Meredith Ringel Morris. 2020. Toward fairness in AI for people with disabilities: A research roadmap. *ACM SIGACCESS Accessibility and Computing* 125 (2020), 1–1.

[12] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 771–787.

[13] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–16.

[14] Sungsoo Ray Hong, Jessica Hullman, and Enrico Bertini. 2020. Human Factors in Model Interpretability: Industry Practices, Challenges, and Needs. In *Proceedings of the ACM on Human-Computer Interaction CSCW*, Vol. 4. 1–26.

[15] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in Neural Information Processing Systems* 33 (2020), 2611–2624.

[16] Yubo Kou and Xinning Gui. 2021. Flag and Flaggability in Automated Moderation: The Case of Reporting Toxic Behavior in an Online Game Community. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan). 1–12.

[17] Toby Jia-Jun Li, Jingya Chen, Haijun Xia, Tom M Mitchell, and Brad A Myers. 2020. Multi-modal repairs of conversational breakdowns in task-oriented dialogs. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 1094–1107.

[18] Neil McDonnell. 2023. The Philosophy of X in XAI. In *Proceedings of the ACM IUI Workshops*, Vol. 2023.

[19] Meta. 2022. BlenderBot 3: An AI Chatbot That Improves Through Conversation. Retrieved 02 23 2023 from https://about.fb.com/news/2022/08/blenderbot-ai-chatbot-improves-through-conversation/

[20] Andrew Sellars. 2016. Defining hate speech. *Berkman Klein Center Research Publication* 2016-20 (2016), 16–48.

[21] Sarah Tan, Julius Adebayo, Kori Inkpen, and Ece Kamar. 2018. Investigating human+ machine complementarity for recidivism predictions. *CoRR* abs/1808.09123 (2018).

[22] Almira Osmanovic Thunström. 2022. We Asked GPT-3 to Write an Academic Paper about Itself—Then We Tried to Get It Published. Retrieved 02 23 2023 from https://www.scientificamerican.com/article/we-asked-gpt-3-to-write-an-academic-paper-about-itself-mdash-then-we-tried-to-get-it-published/

[23] James Vincent. 2022. YouTuber trains AI bot on 4chan's pile o' bile with entirely predictable results. https://www.theverge.com/2022/6/8/23159465/youtuber-ai-bot-pol-gpt-4chan-yannic-kilcher-ethics

[24] Dakuo Wang, Liuping Wang, Zhan Zhang, Ding Wang, Haiyi Zhu, Yvonne Gao, Xiangmin Fan, and Feng Tian. 2021. "Brilliant AI Doctor" in Rural Clinics: Challenges in AI-Powered Clinical Decision Support System Deployment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–18.

[25] Tris Warkentin and Josh Woodward. 2022. Join us in the AI Test Kitchen. Retrieved 02 23 2023 from https://blog.google/technology/ai/join-us-in-the-ai-test-kitchen/

[26] Yong Xie, Dakuo Wang, Pin-Yu Chen, Jinjun Xiong, Sijia Liu, and Sanmi Koyejo. 2022. A Word is Worth A Thousand Dollars: Adversarial Attack on Tweets Fools Stock Prediction. *NAACL* (2022).

[27] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457* (2017).