Responsible Language Technologies: Foreseeing and Mitigating Harms

Su Lin Blodgett*
Microsoft Research
Montreal, Canada
SuLin.Blodgett@microsoft.com

Q. Vera Liao*
Microsoft Research
Montreal, Canada
veraliao@microsoft.com

Alexandra Olteanu*
Microsoft Research
Montreal, Canada
alexandra.olteanu@microsoft.com

Rada Mihalcea
University of Michigan
Ann Arbor, USA
mihalcea@umich.edu

Michael Muller
IBM Research
Boston, USA
michael_muller@us.ibm.com

Morgan Klaus Scheuerman
University of Colorado Boulder
Boulder, USA
mosc1961@colorado.edu

Chenhao Tan
University of Chicago
Chicago, USA
chenhao@uchicago.edu

Qian Yang
Cornell University
Ithaca, USA
qianyang@cornell.edu

ABSTRACT

As increasingly powerful natural language generation, representation, and understanding models are developed, made available and deployed across numerous downstream applications, many researchers and practitioners have warned about possible adverse impacts. Harmful impacts include but are not limited to disparities in quality-of-service, unequal distribution of resources, erasure, stereotyping and misrepresentation of groups and individuals, they might limit people's agency or affect their well-being. Given that language tasks are often complex, open-ended, and incorporated across a diversity of applications; effectively *foreseeing and mitigating such harms has remained an elusive goal.* Towards this goal, Natural Language Processing (NLP) literature has only recently started to engage with human-centered perspectives and methods—that are often central to HCI research. In this panel, we bring together researchers with expertise in both NLP and HCI, as well as in issues that pertain to the fairness, transparency, justice, and ethics of computational systems. Our main goals are to explore 1) how to leverage HCI perspectives and methodologies to help foresee potential harms of language technologies and inform their mitigation, 2) synergies between the HCI and the responsible NLP research that can help build common ground, and 3) complement existing efforts to facilitate conversations between the HCI and NLP communities.

*The first three authors are organizers and moderators, listed in alphabetical order. The rest are panelists listed in alphabetical order

CHI '22 Extended Abstracts, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9156-6/22/04.
https://doi.org/10.1145/3491101.3516502

KEYWORDS

natural language processing, HCI methodologies, Responsible AI, harms measurement

ACM Reference Format:
Su Lin Blodgett*, Q. Vera Liao*, Alexandra Olteanu, Rada Mihalcea, Michael Muller, Morgan Klaus Scheuerman, Chenhao Tan, and Qian Yang. 2022. Responsible Language Technologies: Foreseeing and Mitigating Harms. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI '22 Extended Abstracts), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3491101.3516502

1 PANEL SUMMARY

There is a proliferation of language technologies and applications that are enabled by increasingly powerful natural language generation, representation, and understanding models. While doing so is critical, auditing natural language processing (NLP) systems for possible adverse impacts and outcomes remains an elusive goal. Harmful impacts from language technologies can take many forms, from disparity in quality-of-service and unequal distribution of resources, to erasure, stereotyping and misrepresentation of groups and individuals [4, 5, 18]. Computationally, such harms might occur not only due to what data NLP models are being trained on, but also due to how that data might be embedded and represented by these NLP models and the systems they power, or even due to how NLP systems are often being evaluated [6, 16, 17]. Once NLP applications are deployed, user interactions, appropriation, misuses, and other complex socio-technical dynamics can further exacerbate these harms. In this panel, we start with discussing definitions and examples of harms caused by language technologies.

Developing technologies that benefit people is a core value of the human-computer interaction (HCI) community. Accordingly, the community has contemplated issues related to possible harms, risks, negative consequences, and ethical imperatives for a long time and in many contexts. All the quantitative and qualitative research methods in the HCI methodological toolkit aim to understand the needs of users so technologies can be developed to bring values to them, and some methodologies even have the explicit

goals of preventing or mitigating harms. For example, value sensitive design [9] is an approach to the design of technologies that accounts for the values of stakeholders in a principled manner, to ensure that the technologies created do not violate their ethical values. Participatory design [15, 19] invites stakeholders into the design process to better understand and meet their needs. Speculative design [1] and critical design [3] have long played a role in HCI in challenging the status quo and debating crucial issues that may happen in the future. Feminist design [2], post-colonial design [11], and Queer design [13, 20] have also questioned power relationships that shape technology, as have recent workshops in Human Centered Data Science [14]. This panel will explore *how to leverage HCI methodologies to help foresee or discover potential harms of language technologies, and inform ways to prevent or mitigate the harms*.

Moreover, there are sub-areas of HCI with a particular focus on exploring how technologies and computing research can contribute to diversity and inclusion, social justice, environmental justice, sustainability, and more (e.g. [2, 10]). These HCI sub-areas precede the more recent "AI for good" movement, and have produced rich system contributions, methodological tools, theories, and lessons learned. These areas are growing in recent years, for which CHI introduced a new sub-committee on "Critical Computing, Sustainability, and Social Justice" since 2021 [7]. This panel will also explore *synergies between these lines of HCI research and responsible NLP research to build common grounds towards more inclusive, fair, just and sustainable language technologies*.

Lastly, this panel aims to *facilitate conversations and idea exchange between the HCI and NLP communities*, as a continuity of recent CHI events (e.g., panels [21], workshops [8, 12]) that brought HCI and AI researchers together at CHI conferences. We recognize that there are many open challenges with locating and measuring potential harms that language technologies—and the data they ingest or generate—might surface, exacerbate, or even cause, and want to examine how HCI methodological toolkits can better inform NLP research in this space. We take "foreseeing and mitigating harms" as a central focus to steer the conversations towards building a foundation for human-centered and responsible computing research.

To facilitate these goals, this panel will bring together five panelists from the HCI and NLP communities, with several of them working across both communities. These panelists' expertise spans diverse research topics related to applications of language technologies (e.g., NLP for social good, mitigating behavioral biases, creativity, software engineering) and different perspectives to approaching potential harms of technologies (e.g., human-centered data science, design, AI ethics, social justice). Through interactive question-and-answer with the audience, we aim to explore questions such as:

- What are the potential harms that language technologies can bring? What are some examples of harm?
- How can we define or conceptualize harms for different kinds of language technology applications, and for different kinds of stakeholders?
- How can we foresee potential harms so we can prevent them while developing language technologies?
- How can we identify harms and mitigate them after language technologies are deployed into the real world?
- How can we define and evaluate the success of measures to prevent or mitigate harms?
- How can HCI perspectives and methodologies be used for the above purposes?
- How can we encourage NLP researchers to adopt these HCI methodologies, or better collaborate with HCI researchers?

2 PANEL FORMAT AND PLAN

We plan to adopt a townhall format to maximize the interactions between panelists and the audiences. The panel will start with a short self-introduction from each panelist. We will then open the floor for audiences to ask questions to the panelists. We will mention example questions above to guide the discussions.

Following the format of CHI 2022 conference, we aim to also provide a hybrid experience for the audience. A video conferencing system approved by the conference will be used to support participation of remote audience. To provide a more convenient experience for the remote audiences, we will set up an online system or document for participants to post their questions both during and before the panel. We will also set up an online discussion group for interested audiences and panelists to interact before, during and after the panel. We will invite the audiences to use the online discussion group to inspire a community around the common interests in responsible language technologies and develop future engagement such as co-authored publications, co-organizing workshops, or a special issue of a journal. One of the organizers will focus on moderating the online QA system and discussion group.

We do not expect the panel to incur special logistical needs beyond the common support provided for CHI sessions, such as setting up ways to join for remote audiences and student volunteers on site.

3 PANEL MODERATORS

Given the interdisciplinary focus on HCI, NLP, and Responsible AI, three researchers with backgrounds in all of these areas will organize and moderate the panel. Multiple moderators are also needed to moderate questions and discussions from both the in-person and online audiences. The moderators include:

Su Lin Blodgett: is a postdoctoral researcher at Microsoft Research Montréal, where she works on the ethical and social implications of language technologies. She is a co-organizer of the first and second Bridging Human-Computer Interaction and Natural Language Processing Workshops at NLP conferences.

Q. Vera Liao: is a Principal Researcher at Microsoft Research Montréal. Her current interest is in human-AI interaction, explainable AI, and responsible AI. She serves as the Co-Editor-in-Chief for Springer HCI Series, Editor for CSCW, and on the Editorial Board of ACM Transactions on Interactive Intelligent Systems (TiiS). She actively organizes events that connect the HCI and AI communities, including several workshops and a panel at CHI, IUI and CSCW.

Alexandra Olteanu: is a Principal Researcher at Microsoft Research Montréal in the Fairness, Accountability, Transparency and

Ethics (FATE) group. Her work currently examines practices and assumptions made when evaluating a range of computational systems, with a particular focus on identifying and measuring possible computational harms in language technologies. She sits on the steering committee of the ACM Conference on Fairness, Accountability, and Transparency (FAccT), has served as Tutorial co-chair for AAAI ICWSM 2018, 2020 and ACM FAccT 2019 (where she introduced the "implications tutorials"), and has organized, moderated, and participated in several panels on "FATE concerns in NLP."

4 PANELISTS

To cover a diversity of perspectives, our panelists come from both HCI and NLP communities and include:

Rada Mihalcea: is the Janice M. Jenkins Collegiate Professor of Computer Science and Engineering at the University of Michigan and the Director of the Michigan Artificial Intelligence Lab. Her research interests are in computational linguistics, with a focus on lexical semantics, multilingual natural language processing, and computational social sciences. She currently serves as ACL President. She is the recipient of a Presidential Early Career Award for Scientists and Engineers awarded by President Obama (2009), an ACM Fellow (2019) and a AAAI Fellow (2021). In 2013, she was made an honorary citizen of her hometown of Cluj-Napoca, Romania.

Michael Muller: works as a Research Scientist at IBM Research in Cambridge MA USA (on the traditional, contemporary, and unceded lands of the Wampanoag and Massachusett peoples). His research spans from applications of generative AI to interrogating data science methods for potential social harms. Michael is co-chair of ACM SIGCHI CARES, and serves on the SIGCHI Research Ethics Committee.

Morgan Klaus Scheuerman: is a PhD Student of Information Science at University of Colorado Boulder and a 2021 MSR Research Fellow. His research focuses on the intersection of technical infrastructure and marginalized identities. In particular, he examines how gender and race characteristics are embedded into algorithmic infrastructures and how those permeations influence the entire system. His recent work explores how gender and race classification in computer vision technologies excludes and endangers at-risk individuals.

Chenhao Tan: is an assistant professor of computer science at the University of Chicago, and is also affiliated with the Harris School of Public Policy. He obtained his PhD degree in the Department of Computer Science at Cornell University and bachelor's degrees in computer science and in economics from Tsinghua University. His research interests include natural language processing, human-centered AI, and computational social science. His work has been covered by many news media outlets, such as the New York Times and the Washington Post. He also won an NSF CAREER award, an NSF CRII award, a Salesforce research award, an Amazon research award, a Facebook fellowship, and a Yahoo! Key Scientific Challenges award.

Qian Yang: is an assistant professor in Information Science at Cornell Bowers College of Computing and Information Science. Yang's research investigates emergent HCI design methods and tools that respond to NLP advances (and AI advances broadly), for example, understanding NLP harms in relation to users' use and misuse of the applications, and mitigating such harms via stakeholder participation, service design, and interaction design.

REFERENCES

[1] James Auger. 2013. Speculative design: crafting the speculation. *Digital Creativity* 24, 1 (2013), 11–35.

[2] Shaowen Bardzell. 2010. Feminist HCI: taking stock and outlining an agenda for design. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1301–1310.

[3] Shaowen Bardzell, Jeffrey Bardzell, Jodi Forlizzi, John Zimmerman, and John Antanitis. 2012. Critical design and critical theory: the challenge of designing for provocation. In *Proceedings of the Designing Interactive Systems Conference*. 288–297.

[4] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 610–623.

[5] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 5454–5476.

[6] Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: an inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 1004–1015.

[7] Rob Comber, Shaowen Bardzell, Jeffrey Bardzell, Mike Hazas, and Michael Muller. 2020. Announcing a new CHI subcommittee: critical and sustainable computing. *interactions* 27, 4 (2020), 101–103.

[8] Upol Ehsan, Philipp Wintersberger, Q Vera Liao, Martina Mara, Marc Streit, Sandra Wachter, Andreas Riener, and Mark O Riedl. 2021. Operationalizing Human-Centered Perspectives in Explainable AI. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–6.

[9] Batya Friedman and David G Hendry. 2019. *Value sensitive design: Shaping technology with moral imagination*. Mit Press.

[10] Maria Håkansson and Phoebe Sengers. 2013. Beyond being green: simple living families and ICT. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2725–2734.

[11] Lilly Irani, Janet Vertesi, Paul Dourish, Kavita Philip, and Rebecca E Grinter. 2010. Postcolonial computing: a lens on design and development. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1311–1320.

[12] 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*. 1–8.

[13] Ann Light. 2011. HCI as heterodoxy: Technologies of identity and the queering of interaction with computers. *Interacting with computers* 23, 5 (2011), 430–438.

[14] 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*. 467–473.

[15] Michael J Muller and Sarah Kuhn. 1993. Participatory design. *Commun. ACM* 36, 6 (1993), 24–28.

[16] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data* 2 (2019), 13.

[17] Inioluwa Deborah Raji, Emily Denton, Emily M Bender, Alex Hanna, and Amandalynne Paullada. 2021. AI and the Everything in the Whole Wide World Benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

[18] Morgan Klaus Scheuerman, Jialun Aaron Jiang, Casey Fiesler, and Jed R Brubaker. 2021. A Framework of Severity for Harmful Content Online. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–33.

[19] Douglas Schuler and Aki Namioka. 1993. *Participatory design: Principles and practices*. CRC Press.

[20] Katta Spiel, Os Keyes, Ashley Marie Walker, Michael A DeVito, Jeremy Birnholtz, Emeline Brulé, Ann Light, Pınar Barlas, Jean Hardy, Alex Ahmed, et al. 2019. Queer (ing) HCI: Moving forward in theory and practice. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–4.

[21] Dakuo Wang, Pattie Maes, Xiangshi Ren, Ben Shneiderman, Yuanchun Shi, and Qianying Wang. 2021. Designing AI to Work WITH or FOR People?. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–5.