# Translational AI Center/Dept. of Computer Science Seminar (Spring 2025)


## Nitesh Goyal


**April 18th at 11:00 AM (US Central Time)**

Location and zoom link: https://trac-ai.iastate.edu/event/tras-cs-seminar-nitesh-goyal/

## Impact of Data Annotator Identity on ML Model Outcomes: Unpacking Specialized Rater Pools


## Abstract

Machine learning models are commonly used to detect toxicity in online conversations. These models are trained on datasets annotated by human raters. We explore how raters' self-described identities impact how they annotate toxicity in online comments. We first define the concept of specialized rater pools: rater pools formed based on raters' self-described identities, rather than at random. We formed three such rater pools for this study--specialized rater pools of raters from the U.S. who identify as African American, LGBTQ, and those who identify as neither. Each of these rater pools annotated the same set of comments, which contains many references to these identity groups. We found that rater identity is a statistically significant factor in how raters will annotate toxicity for identity-related annotations. Using preliminary content analysis, we examined the comments with the most disagreement between rater pools and found nuanced differences in the toxicity annotations. Next, we trained models on the annotations from each of the different rater pools, and compared the scores of these models on comments from several test sets. Finally, we discuss how using raters that self-identify with the subjects of comments can create more inclusive machine learning models, and provide more nuanced ratings than those by random raters.


## Short Bio

**Tesh (Nitesh) Goyal (ACM Distinguished Speaker)** leads research at the intersection of Safety and AI, in Responsible AI, Google Research. His work at Google has led to launch of ML based tools like Harassment Manager to empower targets of online harassment, ML based moderation to reduce online toxic content production on platforms like OpenWeb, and multiple NLP based tools that reduce biased sensemaking in criminal justice. He received his MSc in Computer Science from UC, Berkeley and RWTH Aachen, prior to receiving his PhD from Cornell University in Information Science. His research has been supported by German Govt. Fellowship, National Science Foundation, and MacArthur Genius Grant. Frequently collaborating with industry, he has published in top-tier HCI conferences and journals and received two best paper honorable mention awards (CHI, CSCW) and one nomination (ICTD Journal). Tesh has served on the Organization Committee for ACM SIGCHI conferences multiple times and over 10 times as Associate Chair at multiple CHI and CSCW conferences since 2016. Tesh has also been appointed as Adjunct Professor at NYU Computer Science Department.

Translational AI Center

Department of Computer Science

Association for Computing Machinery