I believe that one of the most pressing issues in modern society is the AI value alignment problem: how do we ensure that AI systems pursue goals that benefit humans and are aligned with diverse sociocultural norms? My primary research interest is developing safe and ethical natural language processing (NLP) systems. I am especially interested in:

- NLP Normative Alignment: How can we leverage the breadth of natural language to make intelligent agents and complex natural language systems more culturally and socially aware?

and applying these techniques to:

- NLP for Social Good: Can we use NLP techniques to address issues in social domains?

My previous academic experience has contributed to these interests; in addition to coursework in AI ethics, societal computing, and cognitive science, I have deeply studied linguistics, postmodern philosophy, and critical race theory as a nationally ranked policy debater. My previous research has also guided my interest in the above questions, and during my PhD at the University of Washington I plan to study these questions more deeply and explore interesting new topics in sociocultural NLP and AI.

Building Normatively Aligned AI Systems: The value alignment problem is notably difficult; specifying social norms as reward values is a non-trivial task, and an infinite action space for any open world state leads to infinitely many undesirable outcomes.

In pursuit of value-aligned AI, I joined Professor Mark Riedl's Human-Centered AI Lab at Georgia Tech. I began by investigating transformer model training with a social normativity dataset, becoming more familiar with modern NLP techniques for AI value alignment. This led to my first project: augmenting a principal dataset for AI normativity – the *Moral Stories* dataset. I designed scripts to convert the existing stories into an interactive format for a human study. We prompted participants to describe details about entities and actions within stories, seeking new information that would supplement the existing features of the dataset: necessary preconditions, possible post-conditions, and assumptions made. In doing so, I grew more familiar with challenging assumptions when drawing conclusions situated in cultural norms, which guides my future research.

Next, I investigated fine-tuning language models for storytelling via preference learning. I led human evaluations involving hand-labeling stories based on taxonomies of genres and moral alignments. In a co-authored paper[1], I performed model analysis across various benchmarks, indicating that our fine-tuned model outperformed models 20x larger when evaluating human preference and moral alignment. This work familiarized me with investigating latent abilities of large language models (LLMs), especially pertaining to human alignment and moral character judgement. One of my future research interests is to continue working with implicit cultural norms in LLMs in hopes of understanding and deconstructing emergent biases. I am primarily interested in ways to extract latent normative preferences and train these language models with more robustly-defined normative reward models.

In my senior thesis, titled *Relationship Extraction via Language Models for Normativity Analysis*, I am investigating structured relationship extraction for value alignment. Extracting relationship triples has historically required resource-heavy human annotation. My early work demonstrates that relationship extraction with LLMs, even with implicit world information, is efficient and accurate. Future work involves constructing sequences of knowledge graphs using these triples to evaluate normativity of story characters, and we expect to publish the generated dataset and normativity analysis results. I hope that continuation with this work will contribute to the capability of intelligent agents to analyze the normativity of their own actions, especially pertaining to human interaction.

NLP for Social Good: I became interested in NLP for social good as a research intern at The Ohio State University under Professor Dong Xuan. I worked on an Automatic Speech Recognition (ASR) system for drunk driving detection, and successfully implemented an acoustic language model for speech

[1] Louis Castricato, Alexander Havrilla, Shahbuland Matiana, Michael Pieler, Anbang Ye, **Ian Yang**, Spencer Frazier, and Mark Riedl. Robust Preference Learning for Storytelling via Contrastive Reinforcement Learning. *arXiv preprint arXiv:2210.07792*, 2022. https://arxiv.org/abs/2210.07792

verification. I discovered that a pre-existing speech recognition toolkit was lacking speaker identification capabilities and developed a real-time ID software using clustering techniques with vectorized audio data.

In the summer of 2023, I was an REU student at Carnegie Mellon University, where I am currently a research assistant working with Professor Norman Sadeh on NLP and privacy. During the REU, I refactored a dataset of privacy policies and labels for Android apps and trained machine learning classifiers to predict data collection. I used NLP techniques to clean and vectorize the privacy policies and perform text classification across privacy labels. We noted that syntactically similar policies clustered together in a 2D mapping; this observation helped guide smarter sampling and identify potential noise. These results were presented as a poster following the REU.

Currently, I am evaluating foundation model performance with information extraction over privacy policies, and investigating whether they can perform taxonomic reasoning in automated question-answering tasks. I have found that these LLMs are adept at highly-structured information extraction from legal texts, and a manuscript with these results is in preparation. I also led data processing and analysis from a study where law students annotated privacy policies with answers to common privacy questions, and evaluated these annotations with respect to foundation model performance. In a paper under review at PoPETs[2], I found that many privacy questions require taxonomic inference, and current LLMs are insufficient at drawing these semantic connections. I'm interested in working on future projects incorporating these techniques; I believe that structured information extraction and incorporating reasoning across complex natural language taxonomies in foundation models elicit avenues to develop aligned downstream NLP systems with social norms and requirements.

Future Goals: Ultimately, I hope to be a professor, allowing me to lead research projects and advise students. One of my most fulfilling experiences has been tutoring graduate students enrolled in the Master of Science in Analytics program at Georgia Tech. My proudest moment was when a student – multiple years into his career – asked me how to break into AI research, because he was so deeply interested in the algorithms we were discussing. Through this tutoring experience and by serving as a peer advisor and TA for first-year computer science majors, I have developed a keen interest in teaching and advising, and pursuing a PhD would allow me to pursue both goals.

At the University of Washington, I plan to develop my strict *people-first* approach to language technologies, and I'm primarily interested in working with Professors Yejin Choi and Yulia Tsvetkov on debiasing language models, aligning them with human sociocultural norms, and broader questions investigating the intersections of NLP systems with social values. I am also interested in working with Professor Noah Smith on questions relating to NLP and computational sociolinguistics, such as how to refine language models to be more human-centered and explainable. I would also be interested in working with Professor Natasha Jaques on projects integrating social science and human-centered techniques with machine learning systems writ large. I am also especially drawn to the broader NLP research environment at UW as an institution; by keeping up with the work of current faculty, post-docs, and graduate students at UW and AI2, I am confident that my academic interests and previous research are a strong fit with the group. I believe that UW is an excellent place for me to pursue my PhD.

[2] Abhilasha Ravichander, **Ian Yang**, Rex Chen, Shomir Wilson, Thomas Norton, and Norman Sadeh. 2023. "Incorporating Taxonomic Reasoning and Regulatory Knowledge into Automated Privacy Question Answering". *Under review.*