Personal Statement

As a nationally ranked policy debater and coach, I grew fascinated with the intricacies of language. Guided by social ethics philosophy, I realized that subtle nuances in communication can significantly impact the accessibility of ideas, and learned how difficult it is for a goal (in public policy or otherwise) to be universally beneficial; the tangible effects of cybersecurity education in K-12 schools, immigration policy for asylum seekers, and arms sales to Taiwan cannot be divorced from Black feminist theory from Hortense Spillers, Postcolonial Orientalism from Rey Chow, or the structuralist linguistic theories of Derrida's Deconstruction. While volunteering as the philosophy coach for the Atlanta Urban Debate League, I also began studying computer science at Georgia Tech. I was intrigued by the juxtaposition between the fluidity in ethical theories of language and the rigid logic of computation, and was familiar with philosopher Nick Bostrom's work on the dangers of misaligned AI systems.

As I was exploring potential research labs at Georgia Tech, I was adamant that the projects be not only computationally fascinating, but cognizant of societal prerequisites and implications as well. While taking Professor Mark Riedl's Intro to AI course, I stumbled across his Human-Centered AI (HCAI) Lab's webpage. I was immediately drawn to his work on the human value alignment problem. I appreciated his group's concern about alignment *for whom*, *in what contexts*, and *with respect to which social norms* – familiar variables from my time debating public policies. I approached Dr. Riedl to join the HCAI Lab, and immediately began working with a Ph.D. student to develop a new dataset of textual scenarios based on human sociocultural norms for fine-tuning transformer models, extending the existing *Moral Stories* dataset. I was amazed at how short scenarios could encode complex social norms and provide sufficient training data for a social commonsense model. This experience helped me realize that **natural language processing (NLP) perfectly encapsulated my disparate interests** in computer science, social/ethical philosophy, political science, and linguistics.

I am now committed to imbuing NLP systems with complex, multifaceted concepts from the social sciences, linguistics, and philosophy, with the end goal of developing socially aware language models that stand to benefit everyone, across diverse communities. During my Ph.D., I will collaborate with a network of experts in NLP, psychology, anthropology, philosophy, and public policy, which is both unique and necessary for the sustainable development of NLP for the greater good. After my Ph.D., **I plan to stay in academia as a professor and PI**. Each professor, PI, and mentor that I have worked with has profoundly impacted my maturity as a researcher, and I would be honored to have a similar impact on my own students. I would be incredibly humbled to have the opportunity to pursue a career of research, teaching, and advising, and the National Science Foundation Graduate Research Fellowship would be instrumental in helping me develop as a researcher, scientist, teacher, and mentor in pursuit of these goals.

Intellectual Merit: NLP is a fundamentally interdisciplinary field, and I am hesitant that simply training larger, more powerful foundation models is sustainable. My research focus is to integrate multidimensional concepts from the humanities with traditional computation to develop socially aware language technologies that are more socioculturally aligned. My formal academic training in computer science and math, along with my background in social identity philosophy, political science, ethics, and linguistics, makes me uniquely comfortable with navigating between these domains.

I have continued working with the HCAI Lab to further my goal of building socially aware AI agents. While investigating fine-tuning large language models (LLMs) with preference learning, I led human evaluation where 25 annotators labeled 44 stories based on genres and moral alignments, collaborating with EleutherAI – an open-source, grass-roots AI research group. I performed analysis for GPT-J, GPT-NeoX, and GPT-3, indicating that our fine-tuned model outperformed models 20x larger for human preferences and moral alignment. **This work led to a co-authored paper** and familiarized me with investigating latent value alignment abilities of LLMs, continuing to serve as inspiration for my current alignment research. I learned how to conduct careful and thorough human evaluation of LLM performance and became comfortable with collaborations across multiple labs and organizations, equipping me with the teamwork skills to tackle large-scale research projects.

Following a Ph.D. student's Leave of Absence, I took charge of the remaining work to complete my senior thesis and **produce a second-author paper currently under review for COLING 2025**. My

thesis focused on how we may use LLM-generated knowledge graphs (KGs) to *isolate* character actions as triples ([*entity, action, entity*]) and whether LLMs can detect if actions uphold or violate social norms. I developed a pipeline that collected movie scripts, parsed the scenes and characters, generated KGs from triples, and systematically prompted the LLMs with KGs for character analysis. ChatGPT achieved an F1-score of 0.677 when classifying 248 characters as socially aligned or mis-aligned, indicating that more work is necessary to integrate world knowledge about sociocultural norms. The COLING submission expands on automatic generation of KGs, where we developed a 5-step global-local prompting strategy to encode precondition and effect information with action chaining, allowing fine-tuned T5 models to act as world models in downstream reinforcement learning (RL) tasks. These projects have led to a fascination with understanding what internal representations LLMs learn about the physical and metaphysical worlds, and how we may build world states that steer these internal world models. I am eager to continue research into LLM world models, and especially the latent human biases that are inadvertently encoded.

While working on LLM alignment and RL with Dr. Riedl, I also grew curious about social domains in which we can *deploy* advances in NLP to benefit end-users. In the summer of 2023, **I was selected from over 1,000 applicants** to participate in the NSF-funded Research Experience for Undergraduates in Software Engineering (REUSE) program at Carnegie Mellon University, where I worked with Dr. Norman Sadeh on usable privacy policies in pursuit of this goal. I built a dataset of 17,555 privacy policies and labels for Android apps and trained logistic regression classifiers and BERT-, GPT-, and Llama-based LLMs to predict collection of 38 different data types to identify data disclosure issues. Interestingly, we noted that syntactically similar policies clustered together in a 2D latent space; this observation guided smarter sampling and identified potential noise. An ensemble model achieved an F1-score of 0.606, indicating that many Android Apps have *potential compliance issues* with data collection disclosure requirements, ultimately harming end-users. **These results led to a poster presentation** and **second-authored paper in the Springer Computing Journal.** This work was my first experience with how NLP can directly address human shortcomings in our social spheres. I learned that our architectural/theoretical choices when building language technologies can have direct effects on end users. I now strive to be constantly aware of *who* is affected by my research, in *what ways*.

In continued collaboration with Dr. Sadeh and his group, I evaluated whether LLMs could reason about complex privacy legalese in the context of taxonomic information and regulatory requirements. I worked with Tom Norton, a law professor at Fordham University, to determine if LLMs could correctly interpret missing information in privacy policies in the context of GDPR and CCPA disclosure requirements, finding that ChatGPT was correct in *just 5.36% of cases*. I also examined whether LLMs could be helpful question-answering assistants for everyday cybersecurity questions. I conducted a human annotator study that found that ChatGPT 4 and Llama 2 could provide accurate answers, but not answers that were motivating to users, understandable by diverse groups of users, or actionable for users from varied backgrounds. I also led an evaluation that found that careful prompt engineering could significantly improve performance for these metrics by *as much as 52.02%*. These projects resulted in **two first-author papers to be presented at WISE 2024**, revealing incredible deficiencies in LLMs that must be addressed before adopting widescale use in more pressing social domains. I learned the importance of legal precision in NLP research and look forward to continuing to work with law and social science experts to further investigate shortcomings with LLMs in other user-facing domains. Preparing these publications especially helped me develop the written communication skills necessary to relay results from my research to diverse groups of people from various technical backgrounds.

My previous work at Georgia Tech and CMU revealed gaps in LLM capabilities, and I was eager to explore theoretical approaches to overcome these deficiencies in existing alignment pipelines. To do so, I have been collaborating with Professors Mingyi Hong and Dongyeop Kang at The University of Minnesota to incorporate multi-dimensional human preference data into training dense reward models for reinforcement learning from human feedback (RLHF). I am investigating reward shaping for the LIME (Local Interpretable Model-agnostic Explanations), SHAP (SHapley Addictive exPlanations), and integrated gradients interpretability measures. I developed an RLHF pipeline that distributes the interpretability values as token-level reward signals that augment the reward assigned to each token by an

underlying reward model, finding that SHAP is a candidate metric that could help reveal *which tokens* influence LLM generations the most *for which particular task* (average reward = 2.56, while average reward for a naïve Llama model = 0.45 for the helpfulness/harmfulness generation task). This is especially important as in open RL problems there may be an infinite number of undesirable actions, so defining reward models is incredibly difficult. With dense, multidimensional signals, we are looking to steer the behavior of these LLMs more robustly. Though incorporating this complex data into the RLHF pipeline is an unsolved challenge, it is paramount for aligning LLMs with multifaceted human sociocultural norms. With the prevalence of LLMs in the zeitgeist of today, controllable, interpretable, and socially aware generation is of the utmost importance.

Broader Impacts: The human value alignment problem is often considered one of the most pressing issues when deploying AI for widescale adoption. My research focuses on how we may develop more robust computational techniques to make NLP systems more accessible for all communities. I am eager to collaborate with researchers across CS, the social sciences, and the humanities to further the research in developing safe AI for all. I was especially proud when I discovered that **the paper that I had co-authored with EleutherAI was one of the main papers that led to the founding of CarperAI**, an AI research team focused on democratization of open-source human preference learning.

In addition to my love of research, I am committed to expanding access to computer science and research opportunities for all, especially those from underserved/underprivileged backgrounds. One of my most fulfilling experiences in college was serving as a tutor and mentor for over 1000 students in the Online Master's in Analytics program at Georgia Tech for two years with Dr. Richard Vuduc. The tutoring program was designed for master's students who did not previously have access to formal computer science education. I hosted hourly sessions four times a week for students from across the world. My proudest moment was when an older student who felt stuck in his career asked me how to break into AI research because he was so deeply interested in the ML algorithms we were discussing. He has since joined an AI lab at UC Santa Cruz as a lab technician.

I also had the privilege of directly mentoring 55 freshman computer science students through the College of Computing Peer Mentorship program, working with the Director of Computing Engagement. The Peer Mentorship program was founded knowing that many students come from underserved communities for computer science. I guided the students through their transition into undergraduate life and helped build schedules, create resumes, and manage stress and mental health. It was incredibly rewarding watching these students begin their college experience unsure about their educational backgrounds and grow into true scholars. I have kept in contact with many of these students even years later, and am excited for similar mentorship roles during my Ph.D.

While working with Dr. Sadeh and his research group this past summer, I also had the invaluable opportunity to mentor a new REU student for her own project – my first time acting as a direct research mentor. I guided her through project onboarding and continued direct collaboration throughout the summer via weekly meetings, culminating in an accepted conference paper. As an eventual Ph.D. student and professor, **I am especially looking forward to mentoring more REU students**, as I strongly believe that an REU is a uniquely impactful undergraduate experience, especially for under-represented students.

During my final two years of college at Georgia Tech, I was chosen to be one of 28 VOICE peer educators. VOICE is Georgia Tech's sexual and relationship violence prevention and survivor support initiative, and the peer education program exists to spread domestic violence awareness and prevention resources. All the peer educators shared the same ideal: no one should be worried about their safety while getting an education. I attended biweekly seminars where we discussed issues related to sexual violence and developed outreach strategies and presentation ideas to educate others on prevention strategies and techniques. I encountered many tough circumstances as a peer educator and developed confidence in navigating extremely sensitive situations.

Throughout my time as a peer mentor, tutor, and educator, I have grown acutely aware of systemic issues and disparities that have profound impacts on someone's educational experience. I hope to use what I have learned in pursuit of equality in computer science, and the GRFP would be instrumental in allowing me to not only pursue research, but mentorship and outreach as well.