Graduate Research Plan

Introduction: Recent improvements in natural language processing (NLP) have led to the trend of training models with naïve performance benchmarks as the end goal, forgetting that humans are at the crux of language. An ongoing challenge is robust alignment of large language models (LLMs) with complex human values. The vast text data used to train state-of-the-art models is not insulated from human biases and current alignment methods lack the necessary guidance from the social sciences and humanities to be viable in the long term, often only concealing biases rather than eliminating them [1]. My research will draw from implicit bias mitigation techniques from psychology and anthropology and ethical theories from philosophy to inform bias evaluation for LLMs and investigate changes to LLM fine-tuning that will support multidimensional human values data. I am particularly interested in: ***RQ1: How can we systematically analyze latent social biases in LLMs?*** and ***RQ2: How can we steer LLMs to act in ways that are aligned with diverse, intersectional human sociocultural norms?*** This work has the potential to both reveal and address incredible social deficiencies in our current SOTA models.

Background: The human normative alignment problem aims to develop AI systems that act in accordance with human values and often involves reinforcement learning from human feedback (RLHF), using human preference data as rewards to fine-tune LLMs using proximal policy optimization (PPO). This process has been effective at generating outputs that are more desirable for humans, with perhaps the most well-known success of RLHF being the development of ChatGPT. Even with RLHF and techniques such as direct preference optimization, LLMs remain imperfect and often act undesirably [1, 2, 3, 4].

Proposal: My objective is to develop a framework to study and develop socially aware NLP models that are aligned with diverse sociocultural norms. I propose two avenues of work to address the ***RQs***:

Approach 1: **Isolating Learned Human Norms in LLM World Models** LLMs reproduce and even amplify biases in training data, inheriting prejudices related to gender, race, disability, and mental health status [2]. Using *probing*, which leverages linear regression models trained on network activations, recent work demonstrates that LLMs learn an internal model of the world, including spatial and temporal dimensions [5]. Linear probes may also be applied to situational data to uncover an entity's dynamic properties, or even relationships with other entities [6], yet no existing research has applied linear probes to human sociocultural norms. **The first component of my research will involve probing internal world models using a dataset of *social world states***. My senior thesis briefly explored this problem by prompting LLMs with sequences of knowledge graphs (KGs) to evaluate entity actions. The datasets I will build from are *JerichoWorld* [7], a dataset of text world states that, while influential, is lacking in size and scope, and *Moral Stories* [8], a crowd-sourced dataset of branching narratives for social reasoning. I will automatically generate KGs [9] to collect additional world states and model sequences of actions *in situ* to augment *JerichoWorld* and use the crowdsourcing platform Prolific to annotate the world states in the branched human norm narratives style from *Moral Stories*. I will then train linear regression probes with this dataset to examine the presence and representations of latent *social norms and biases* in internal LLM world models. I will craft prompts and develop strict human annotation guidelines based on the Prejudice-Habit Breaking framework – one of the only unconscious bias training frameworks with proven longitudinal results [10] – and ensure annotators from diverse demographic backgrounds annotate the same world states to gain insights into *intersectional* latent sociocultural values. To determine the effectiveness of my probing strategy, I will use the LLM Implicit Bias Test [3], developed from the human Implicit Association Test, to systematize the evaluation of learned biases revealed by the probes, thus unifying the probing and evaluation stages.

Approach 2: **Developing Dense, Multidimensional Normative Reward Signals for RLHF** One of the main limitations of RLHF is that reward models necessarily need to output *scalar* rewards for RL finetuning. This evokes an obvious issue: are scalar rewards nuanced enough for effective alignment? Recently, reward-shaping has emerged to artificially "densify" reward signals and multi-objective RL has been used to build multidimensional reward models that are selected using a Mixture-of-Experts (MoE) approach or via a linear combination of signals [11]. Even still, research is lacking in incorporating reward signals that are both *dense* and *multidimensional*. **The second component of my research will involve redefining how multidimensional rewards are distributed, and thus how reward models are**

incorporated into RLHF fine-tuning. My previous research indicates that interpretability measures such as LIME (Local Interpretable Model-agnostic Explanations), SHAP (SHapley Addictive exPlanations), and integrated gradients can provide reward signals in parallel with fine-tuned reward models to identify important tokens in generations *for specific alignment tasks*, such as helpful-/harmful-ness. These interpretability measures are unique because they are naturally dense; SHAP assigns interpretability at the *token-level*, which is distributed across all tokens to shape the underlying reward model signal. I will apply similar reward distribution techniques at the *token-level* using multi-dimensional reward signals to build more robust multi-objective reward signals. As proof of concept, I will use the *HummingBird* dataset [12] which includes text data annotated by humans for politeness, sentiment, anger, disgust, fear, joy, and sadness. I will then work with anthropologists at my graduate school and recruit identity ethics philosophers to build a dataset of *entity actions* that are labeled by human annotators in accordance with *cross-cultural* human norms (cross-cultural studies is a pressing limitation of current alignment research [4]). By applying multi-objective RL and reward-shaping techniques to this new *cross-cultural* dataset, I will develop the first RLHF pipeline that inherently focuses on *intersectional normative alignment*.

Intellectual Merit: This work proposes a framework that investigates, evaluates, and mitigates intersectional biases in LLMs. Successful implementation of this alignment framework would contribute to a new perspective for NLP research. Instead of training larger and larger foundation models, or even highly specialized models for one or two specific tasks, the focus would shift towards satisfying suites of social reasoning problems with new techniques to incorporate human values. Having a system to probe LLMs to identify deep-rooted biases gives researchers a way to blueprint internal world models, furthering the understanding of current "black box" models. Incorporating dense, multidimensional reward signals into RLHF will make models far more adept at diverse downstream alignment tasks, overcoming limitations from "aligning to the mean" and will directly influence future LLM development. Much like how humans only change their behavior when directly confronted with their own implicit biases [10], this research has the potential to alter the course of NLP research to integrate the social sciences and humanities by forcing researchers to confront deficiencies in current development pipelines.

Broader Impacts: The human value alignment problem is fundamentally concerned with whether intelligent systems are safe to deploy, for whom, and in what instances. Even SOTA models stand to benefit white men more than any other demographic group. As LLMs pervade all aspects of our daily lives, from chatbots, to writing assistants, and even political trolls on social media, it becomes increasingly important to have a more robust framework to ensure that these models are aligned with human sociocultural norms and values. Successful implementation of this work will lead to more critical deployment of LLMs that will safely benefit diverse *individuals* with *unique sociocultural backgrounds*. LLMs should not be better assistants, tutors, or advisors for white men than for Asian men, Black women, or Queer folk (and should not perpetuate prejudices with faulty generations), and the above alignment framework will be an initial step towards a more equitable future.

References: [1] Gonen, H., & Goldberg, Y. "Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them". NAACL 2019. [2] Field, A., et al. "Examining Risks of Racial Biases in NLP Tools for Child Protective Services". ACM FAccT 2023. [3] Bai, X. et al. "Measuring Implicit Bias in Explicitly Unbiased Large Language Models". arXiv 2024. [4] Choi, M., et al. "Do LLMs Understand Social Knowledge? Evaluating the Sociability of Large Language Models with SocKET Benchmark". ACL 2023. [5] Gurnee, W., & Tegmark, M. "Language Models Represent Space and Time". arXiv 2023. [6] Li, B. et al. "Implicit Representations of Meaning in Neural Language Models". ACL 2021. [7] Ammanabrolu, P., & Riedl, M. O. "Modeling Worlds in Text". NeurIPS 2021. [8] Emelin, D. et al. "Moral Stories: Situated Reasoning about Norms, Intents, Actions, and their Consequences". EMNLP 2021. [9] Ye, H., et al. "Generative Knowledge Graph Construction: A Review". arXiv 2022. [10] Devine, P. et al. "Long-term reduction in implicit race bias: A prejudice habit-breaking intervention". JESP 2012. [11] Wang, H. et al. "Arithmetic Control of LLMs for Diverse User Preferences: Directional Preference Alignment with Multi-Objective Rewards". arXiv 2024. [12] Hayati, S. A., et al. "Does BERT Learn as Humans Perceive? Understanding Linguistic Styles through Lexica". EMNLP 2021.