Research Summary – CRA Outstanding Undergraduate Researcher Award

Eisuke Hirota

Reinforcement Learning with Human Feedback (RLHF) is the primary technique used to align Artificial Intelligence (AI) with human values. RLHF has encountered a fundamental obstacle, however – current preference aggregation methods fall short in considering *pluralistic alignment*, the ability to comprehend diverse perspectives. As such, during my PhD and onwards, I will develop aggregation methods that create the theoretical foundation for pluralistic alignment through three core concepts:

1) **Modality:** Current preference aggregation methods convert voting rules to algebraic aggregation methods; however, voting rules are not applicable to some types of feedback modalities such as language. How can we aggregate such feedback that best represents the population?
2) **Safety:** Control Barrier Functions (CBFs) provide a framework for a user to align embodied AI with her personal safety constraints. How can we aggregate multiple users' CBFs to develop an agent that acts as safely as the population warrants it?
3) **Commonsense:** Inherently, society agrees on some set of features – this is one type of commonsense. Can we extract commonsense from information-rich data such as task-agnostic play data or foundation models to pretrain reward models for RLHF?

My research experience has particularly prepared me to solve these problems. Specifically, I crystallized my expertise in RL through exposure in three labs, built up my research skills, and formed my career goal to become a research scientist and improve government policy in AI regulation.

Construction Robots: Under Professor Chen Feng at NYU, I helped propose a novel RL algorithm that decouples robot localization and planning into two separate models, performing twice as better than state-of-the-art baselines. This method is crucial for construction tasks where robots trained using state-of-the-art baselines cannot simultaneously localize and plan itself. This occurs because the environment is not static – a robot builds its own environment that it must plan around. Thus, by using two separate models, one for localization and one for planning, our robot can construct target structures with higher success. We published this work at ICLR 2023, and I presented the poster and video for this conference.

Similarly, in RLHF, there exists work that compare end-to-end models against decoupled models which can voice individual voices more effectively, but at the cost of computational resources. This has inspired me to focus on aggregation methods that can better represent the population with the fraction of the cost.

Guide Dog for Visually Impaired Users: With Professor Shiqi Zhang at Binghamton University, I helped develop a novel force estimator to estimate the external forces applied onto a quadrupedal robot, achieving up to 80% higher accuracy than the onboard accelerometer. Through the addition of this force estimator, our system accomplishes two things: (1) we feed the estimated forces into the policy, enabling a robust, force-tolerant quadruped controller, and (2) we can classify the direction to which the force is being applied. Through (2) we developed a guide-dog system for the visually impaired, emulating real guide dogs. Our work was accepted at CoRL 2023, and I presented a poster for our project at the SUNY AI Symposium 2023.

Guide dogs safely direct users through the environment – this made me question whether we can achieve safety through RLHF. Currently, RLHF does not explicitly model safety preferences, hence, I believe that the integration of CBFs can tackle this problem.

Reward Learning: Alongside Professor Erdem Bıyık at USC, I developed an active learning approach that easily learns a human's reward function using Bayesian inference. I enabled two features: (1) allow humans to reply with language feedback to our queries, giving us more information to learn the reward function, and (2) a language-integrated information gain equation that actively chooses the next best query. These contributions achieve substantially faster and more optimal convergence of learned reward functions. This research thread is due for submission by mid-November 2024.

Since current RLHF methods integrate massive corpuses of data with many self-bootstrapping steps, there exists room for improvement for faster and more optimal convergence. Through my work with Bayesian inference, I believe that a strong prior can greatly reduce computational cost, henceforth I argue for extracting commonsense for pretraining reward functions.

Future: I plan to pursue a PhD with consideration of pluralistic alignment within RLHF. AI will greatly impact society, both positively and negatively, and hence I hope to lay the theoretical foundation in ensuring that they act according to society's values. I believe science alone is not enough, however, and so I also aim to advance AI regulation and push for policy development.