(Introduction, no header)

Aligning systems to human goals requires vast human-annotated data, and often assumes those alignments are final; yet human goals switch, and as models require more data, updating annotations to reflect updated goals is infeasible.

For this reason,

During my undergraduate career, …

Research Background

Wanted to understand transformer attention, Hao attention variants (global, sliding window, random)

- Outcome

Wanted to understand human alignment, RLHF/PPO with TRL and Yushi

- Outcome

Wanted to understand output model biases, controlling these (preference) given frozen model, Hila work with consistency/sensitivity

- Outcome

ColorGrid?

- Outcome

Future Goals

..

Conclusion

<specific to school>

Drafting Below ///

..

Human-system collaboration is a well-studied field of computer science <this sounds like HCI> / deep learning. Aligning a system to a human's goals is done explicitly in natural language processing (NLP) using human annotations to build human annotation estimators, called "reward models", and done implicitly in reinforcement learning (RL) by simulating cooperative play with populations of other diversely capable agents.

A natural extension of goal alignment is to support multiple human goals concurrently; in NLP, pluralistic alignment supports modeling contradictory goals for different groups of people, and in RL, agents use language models to reason about the state of mind of different agents. Both are explicit modeling solutions, relying on human annotations or on human-curated templates for reasoning.

Yet, learning to estimate human goals that change over time is understudied. RL cooperative agents excel at a specific task, yet adapting them to a different goal is challenging. Language models provide general cross-task reasoning but are not experts and require explicitly encoding goal changes. In the way humans attune to implicit behavioral cues to detect goal shifts, I believe systems will better collaborators when trained to do the same. Like AlexNet transitioning from human-engineered to learned features, implicitly learned policies over behavior cues will perform better than explicit human annotated behavior. An ancillary benefit is that training agents in simulation does not require slower, costly human annotation.

My research agenda is toward building expert collaboration systems by studying how models represent dynamically switching goals, how to learn implicitly from observed collaborator behavior, and how simulation-generated data reduces annotation time and cost. Together, these directions create

1) Goal switching
2) Implicit vs explicit
3) Feasible via simulation

, and learning to do so implicitly by simulation, remains elusive. In real human-to-human collaboration,

Implicitly learning goals will be more effective than explicitly learning them.

Parallel of image processing, human crafted features vs learned features

Data challenge

However, in the real world it can be tedious or challenging to

Yet another natural extension is depth vs breadth: supporting the recognition of changing goals in a game

In natural language processing, goal alignment is explicitly enforced by modeling human annotations (RLHF, DPO, KTO, CLAIR). In reinforcement learning, goal alignment is implicitly learned when cooperation performs empirically better in a simulation with other humans or agents (StarCraft, Capture the Flag, SelfPlay, CoPlay DeepMind).

Humans are diverse, and understanding/modeling different/contradictory goals is one facet to build systems that better interface with this diversity; progress in this direction includes explicitly modeling different user annotations (Plurality Survey by Taylor, Yejin) and explicitly reasoning over other agents' state of mind using a large language model (Sotopia, Concordia).

However, an understudied facet of human collaboration is when a human's goal changes {truly understudied? Sotopia/Concordia? Explicit vs implicit goal communication/modeling, which is different than implicit/explicit human signal}

<threads in ChatGPT idea>

During my undergraduate career, I

- Not okay to say "explored", unify around themes

Research Background

..

Wanted to understand transformer attention, Hao attention variants (global, sliding window, random)

Wanted to understand human alignment, RLHF/PPO with TRL and Yushi

Wanted to understand output model biases, controlling these (preference) given frozen model, Hila work with consistency/sensitivity

ColorGrid paper

Future Goals

..

Conclusion

<specific to school>

Self AR

MS to keep discovering, research vs industry; perhaps higher level observations LLM > arch, reward modeling; broader classes to discover the “it” pulling me to study. Human goal inference