Most RLHF algorithms assume an underexamined \textit{partial return} model of human preference. We previously found that another model based on regret better describes human preferences.

What are the consequences of this mistaken assumption?

Learning Optimal Advantage from Preferences and Mistaking it for Reward

W. Bradley Knox\textsuperscript{1,2} Stephane Hatgis-Kessell\textsuperscript{1} Sigurdur Orn Adalgeirsson\textsuperscript{1} Serena Booth\textsuperscript{3} Anca Dragan\textsuperscript{4} Peter Stone\textsuperscript{1,5} Scott Niekum\textsuperscript{6}

First two authors contributed equally

\textsuperscript{1}The University of Texas at Austin
\textsuperscript{2}Google Research
\textsuperscript{3}MIT CSAIL
\textsuperscript{4}UC Berkeley
\textsuperscript{5}Sony AI
\textsuperscript{6}University of Massachusetts Amherst

Background

- **Which fits your preferences?**
- **Which shows better behavior?**

Experiments in 30+ gridworld MDPs

- **When $A^*_p$ is exactly known**
  - Optimal policies are preserved.
  - An underspecification issue is resolved where choice of discount factor ($\gamma$) can be impactful yet arbitrary.
  - Reward is highly shaped, effectively setting $\phi(s) = V^*(s)$ as recommended by Ng et. al. (1999).
  - Since $\arg\max A^*_p$ creates an optimal policy, using $A^*_p$ as reward wastes computation and environment sampling.

- **When $A^*_p$ is approximated as $A^*_g$**
  - Adding transitions from absorbing state to early-terminating segments ameliorates this issue.
  - Including segments with transitions from absorbing state encourages $\max A^*_g = 0$.
  - Arbitrary bias towards or against termination determines performance differences:

<table>
<thead>
<tr>
<th>Condition</th>
<th>$\pi^*_g$ terminates</th>
<th>$\pi^*_g$ does not terminate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max local partial return &gt; 0</td>
<td>$Q^<em>_c$, $Q^</em>_n$</td>
<td>$\hat{A}$, $\hat{A}$</td>
</tr>
<tr>
<td>Max local partial return &lt; 0</td>
<td>$Q^<em>_c$, $Q^</em>_n$</td>
<td>$\hat{A}$, $\hat{A}$</td>
</tr>
</tbody>
</table>

- When adding absorbing transitions, reward is also highly shaped with the approximation error of $\hat{A}^*_g$.

General results: using optimal advantage as reward

- Shaping results may explain why the partial return preference model often performs well.
- Revealed large pitfall and amelioration by including absorbing states in early-terminating segments.
- Offers a simpler reframing of the main method for fine-tuning LLMs with RLHF.