LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

ONE MUNICH Strategy Forum
Next Generation Human-Centered Robotics

Munich Center for Machine Learning

SPONSORED BY THE
Federal Ministry of Education and Research

DAAD Zuse Schools
Konrad Zuse Schools of Excellence in Artificial Intelligence

On the Challenges and Practices of Reinforcement Learning from Real Human Feedback

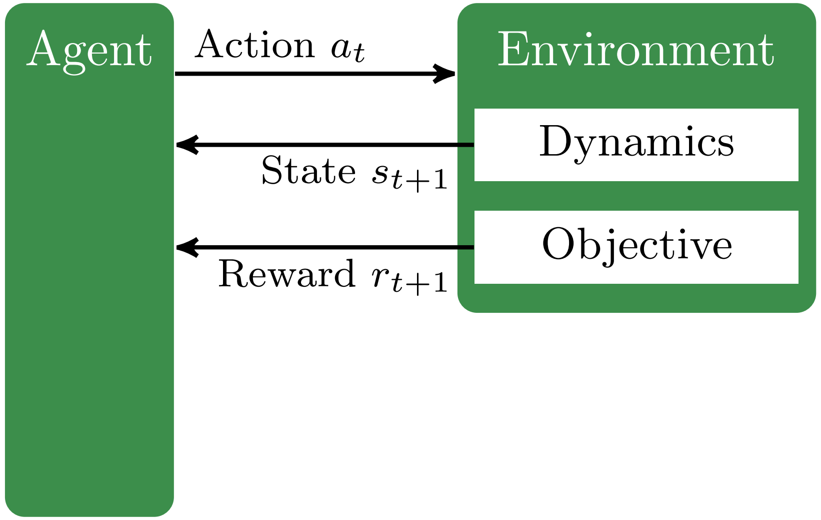
Timo Kaufmann, Sarah Ball, Jacob Beck,
Eyke Hüllermeier, and Frauke Kreuter

LMU Munich

Turin, 2023-09-22

Reinforcement Learning

Learning from rewarded interaction with an environment.

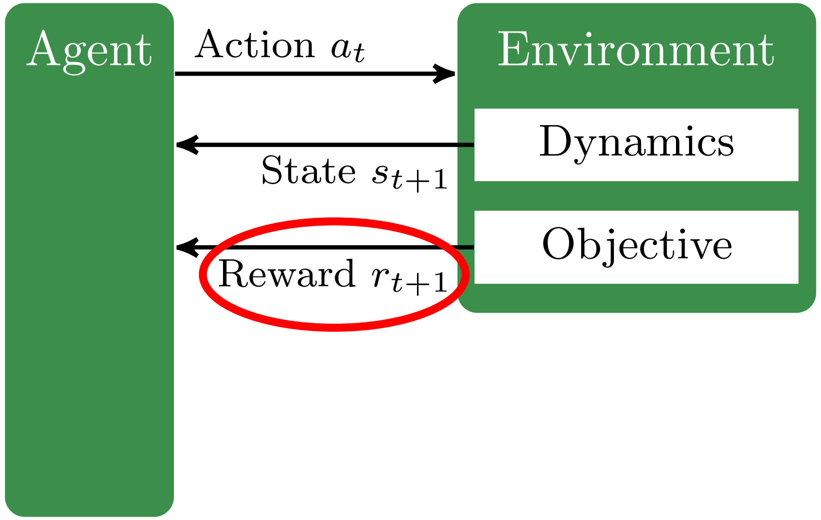


Goal: Find policy π that maximizes

$$J(\pi, s_0) = \mathbb{E}_{\pi, s_0}\left[\sum_{t=0}^{T} \gamma^t r_t\right]$$

Reinforcement Learning

Learning from rewarded interaction with an environment.

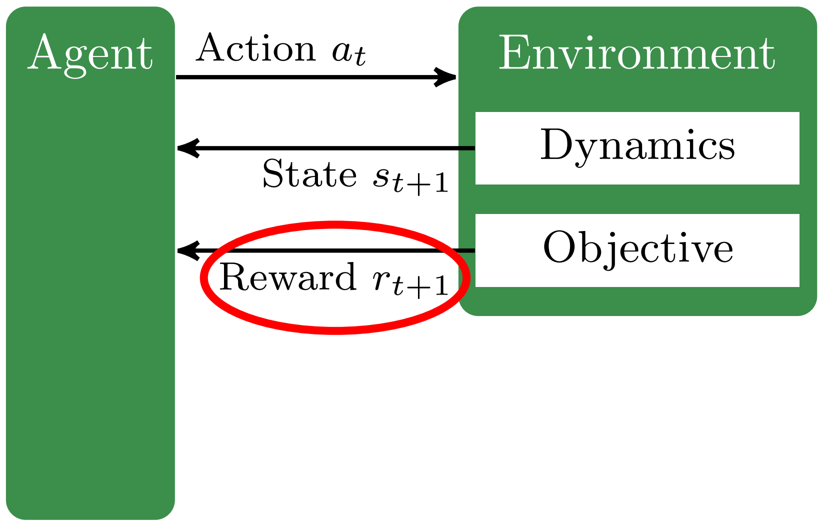


Goal: Find policy π that maximizes

$$J(\pi, s_0) = \mathbb{E}_{\pi, s_0}\left[\sum_{t=0}^{T} \gamma^t r_t\right]$$

Reinforcement Learning

Learning from rewarded interaction with an environment.

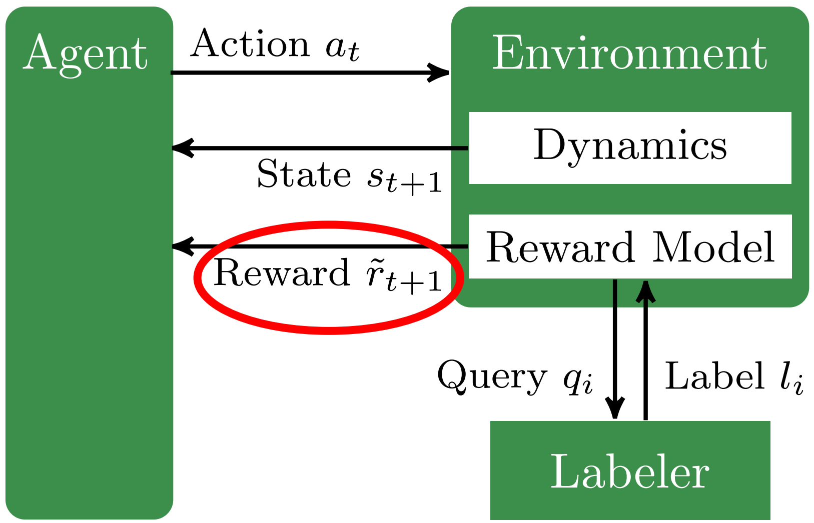


Goal: Find policy π that maximizes

$$J(\pi, s_0) = \mathbb{E}_{\pi, s_0}\left[\sum_{t=0}^{T} \gamma^t r_t\right]$$

From Human Feedback

Defining rewards that induce desired behavior is challenging → **RLHF**



Feedback on **trajectories**

$$\tau_i = (s_0, a_0, s_1, a_1, \ldots, s_n, a_n)$$

Note: This slide contains videos, which have been replaced by single frames for the PDF export.

Pairwise Comparison Example

τ_1

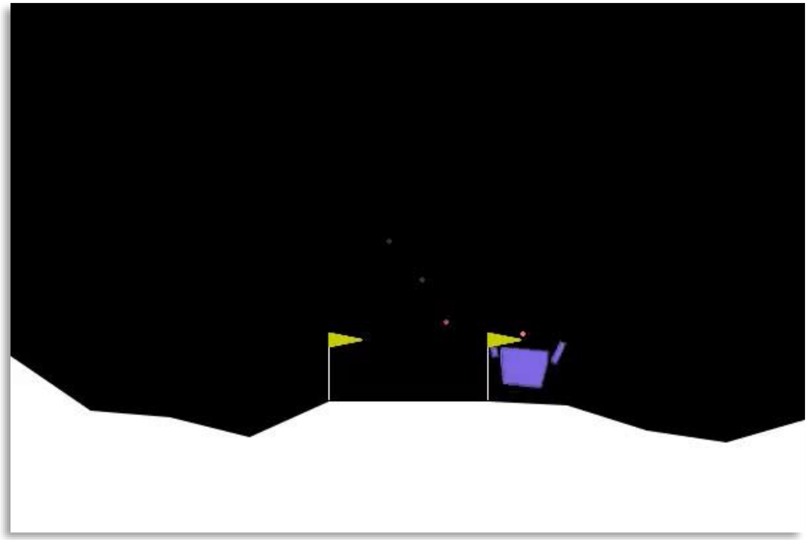
τ_2

[1] Jeon et al., 2020, NeurIPS

Note: This slide contains videos, which have been replaced by single frames for the PDF export.

Pairwise Comparison Example

τ_1



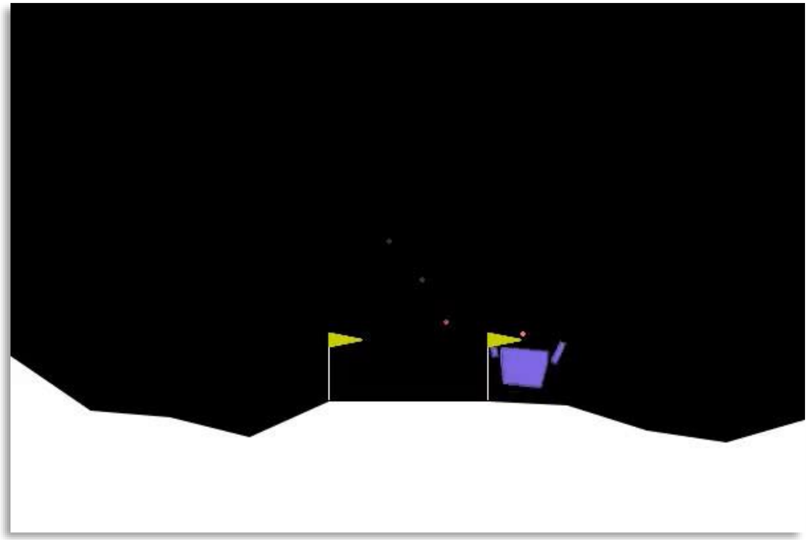
τ_2

[1] Jeon et al., 2020, NeurIPS

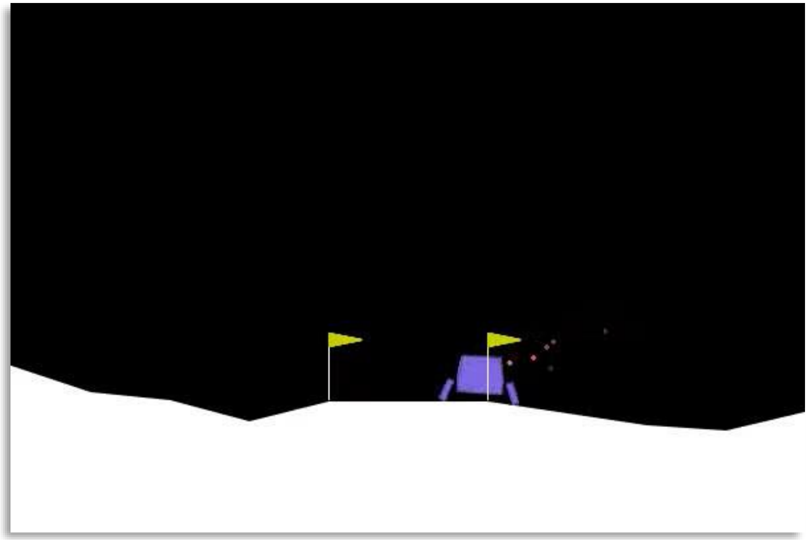
Note: This slide contains videos, which have been replaced by single frames for the PDF export.

Pairwise Comparison Example

τ_1



τ_2

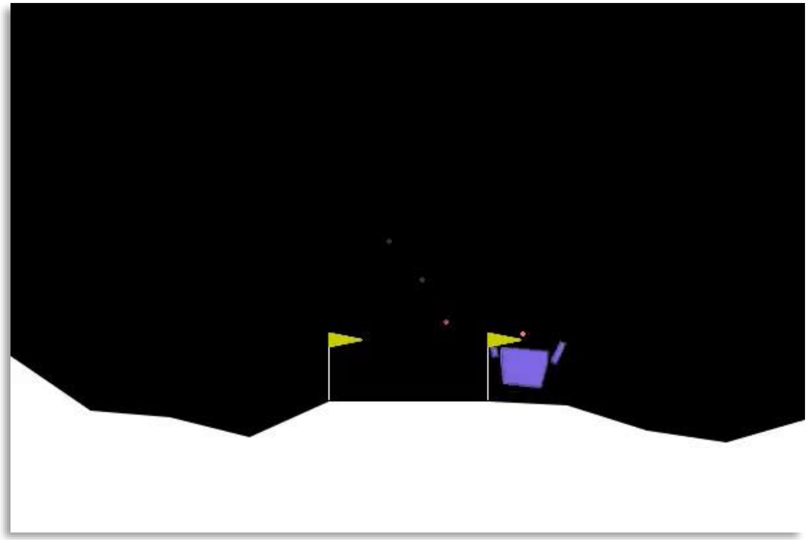


[1] Jeon et al., 2020, NeurIPS

Note: This slide contains videos, which have been replaced by single frames for the PDF export.

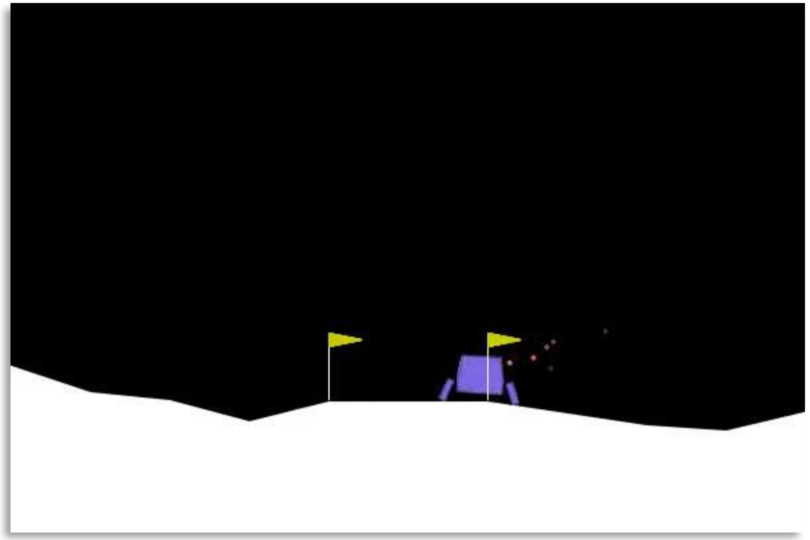
Pairwise Comparison Example

τ_1



$\prec$

τ_2

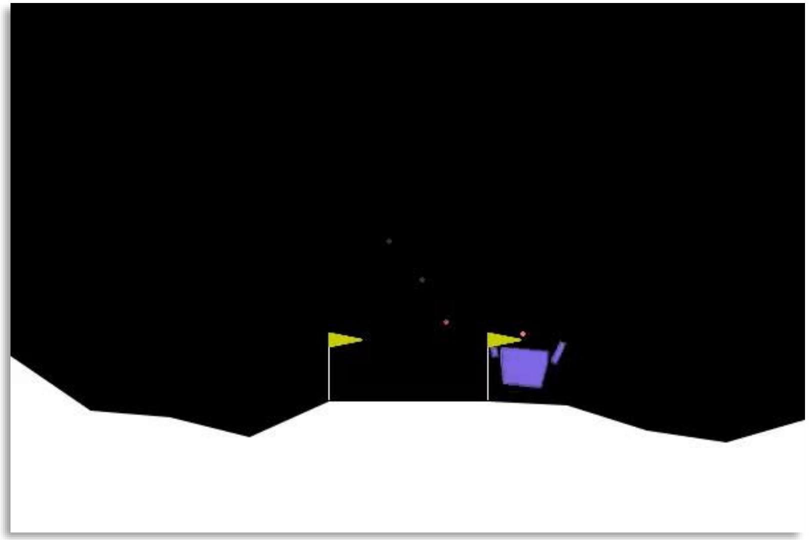


[1] Jeon et al., 2020, NeurIPS

Note: This slide contains videos, which have been replaced by single frames for the PDF export.

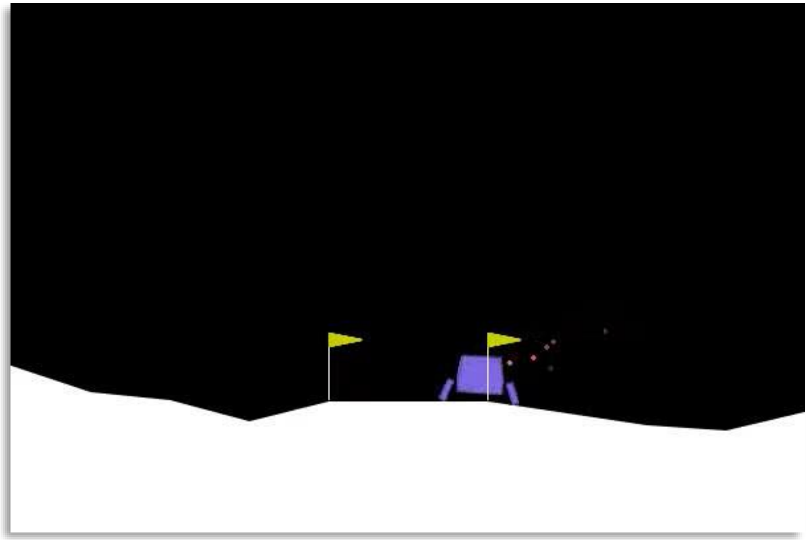
Pairwise Comparison Example

τ_1



$\prec$

τ_2

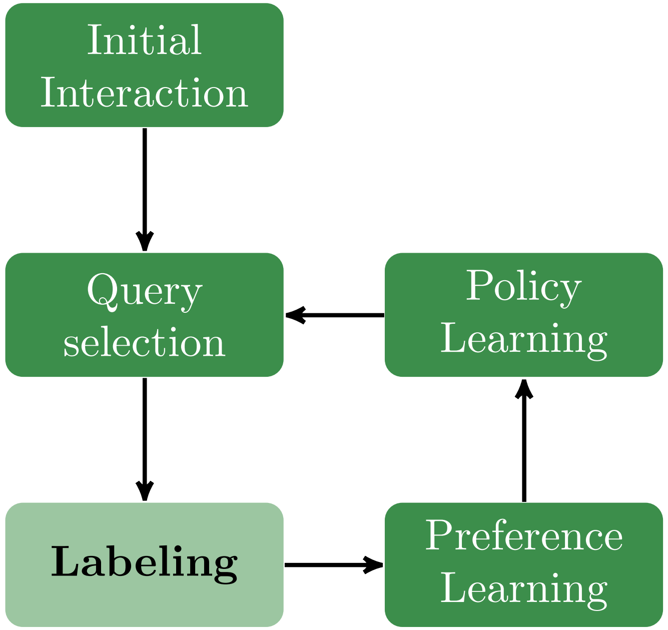


Assumption: Labeler makes reward-rational[1] choice.

$$\mathbb{P}(\tau_1 \succ \tau_2) = \frac{\exp R(\tau_1)}{\exp(R(\tau_2)) + \exp(R(\tau_1))}$$

[1] Jeon et al., 2020, NeurIPS

Labeling is Important



- Real human feedback is inconvenient.
- Researchers often synthesize feedback for evaluation.
- Our argument: This is not enough!

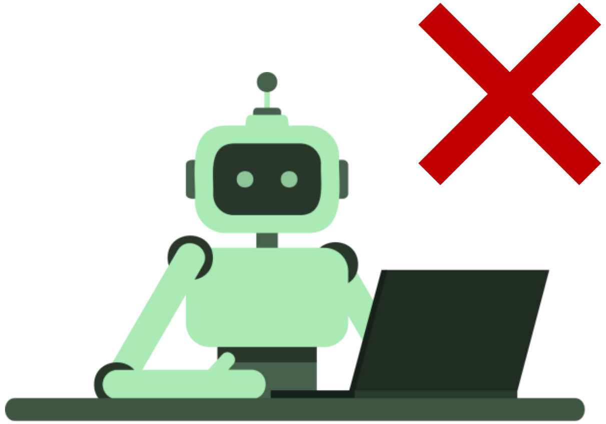


Image: Freepik.com

Challenges of Real Human Feedback

- Response biases, inconsistent behavior
  - Acquiescence bias
  - Primacy/recency effects
- Unobserved factors
  - Motivation
  - Distraction
- Disagreements
  - Intra-labeler (fatigue, experience, …)
  - Inter-labeler
  - Researcher-labeler (misunderstandings)

Images: Leonardo.Ai

Challenges of Real Human Feedback

- Response biases, inconsistent behavior
  - Acquiescence bias
  - Primacy/recency effects
- Unobserved factors
  - Motivation
  - Distraction
- Disagreements
  - Intra-labeler (fatigue, experience, …)
  - Inter-labeler
  - Researcher-labeler (misunderstandings)

Images: Leonardo.Ai

Challenges of Real Human Feedback

- Response biases, inconsistent behavior
  - Acquiescence bias
  - Primacy/recency effects
- Unobserved factors
  - Motivation
  - Distraction
- Disagreements
  - Intra-labeler (fatigue, experience, …)
  - Inter-labeler
  - Researcher-labeler (misunderstandings)

Images: Leonardo.Ai

Challenges of Real Human Feedback

- Response biases, inconsistent behavior
  - Acquiescence bias
  - Primacy/recency effects
- Unobserved factors
  - Motivation
  - Distraction
- Disagreements
  - Intra-labeler (fatigue, experience, …)
  - Inter-labeler
  - Researcher-labeler (misunderstandings)

Images: Leonardo.Ai

Opportunities of Real Human Feedback

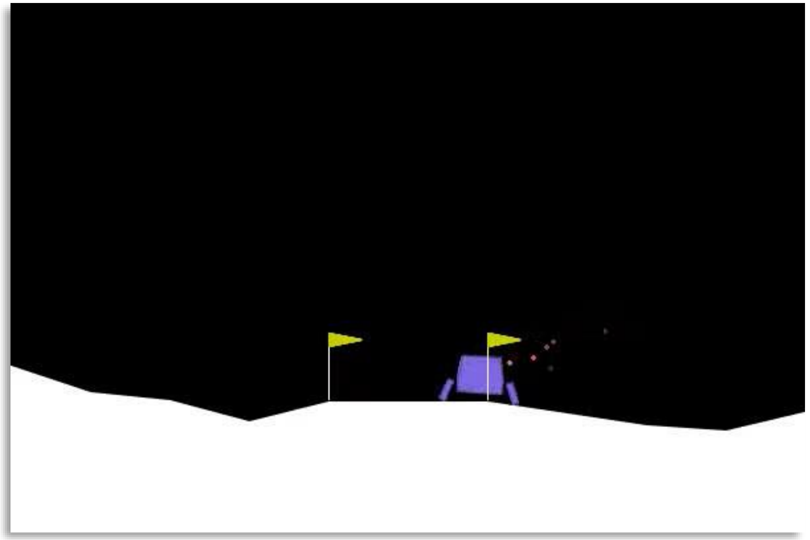
- Optimize the labeling task
  - Goal: obtain more feedback for the same amount of human time
  - Extend or replace comparison queries (e.g. explanations, more response options, long interactions)

[2] Wilde et al., 2022, IJRR

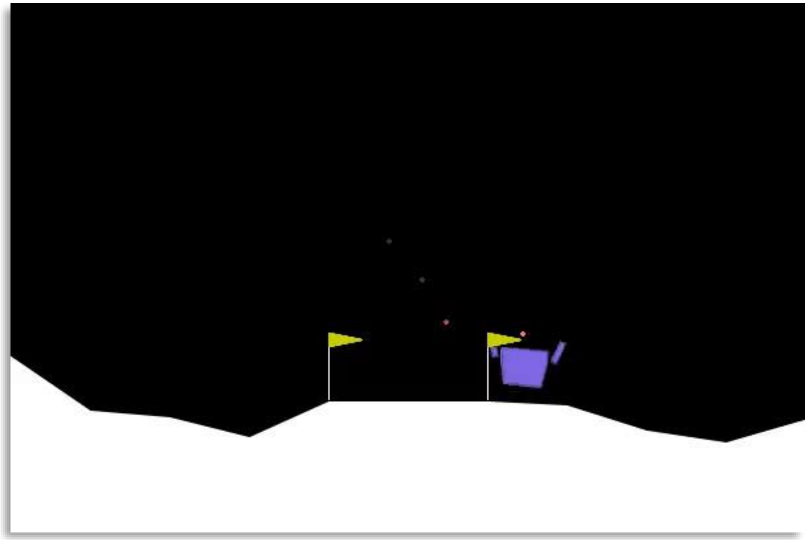
Opportunities of Real Human Feedback

- Optimize the labeling task
  - Goal: obtain more feedback for the same amount of human time
  - Extend or replace comparison queries (e.g. explanations, more response options, long interactions)

τ_1



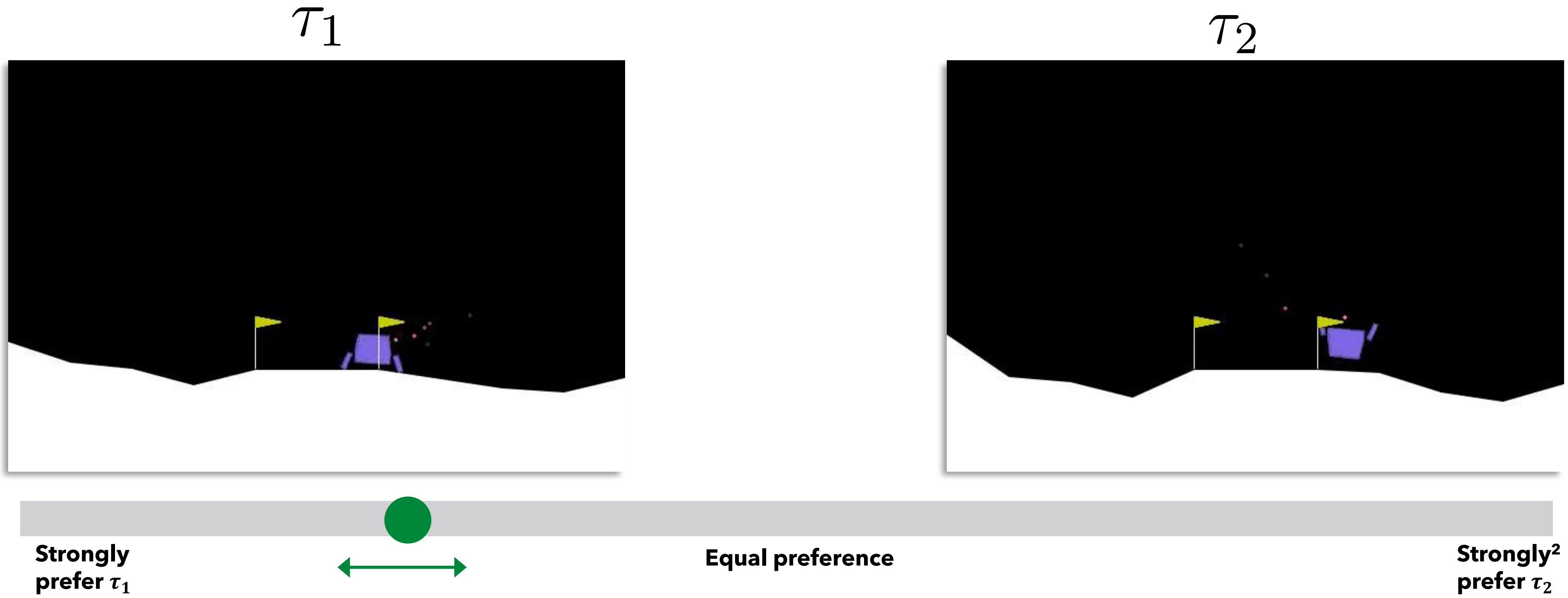
τ_2



[2] Wilde et al., 2022, IJRR

Opportunities of Real Human Feedback

- Optimize the labeling task
  - Goal: obtain more feedback for the same amount of human time
  - Extend or replace comparison queries (e.g. explanations, more response options, long interactions)



τ_1

τ_2

Strongly prefer τ_1

Equal preference

Strongly[2] prefer τ_2

[2] Wilde et al., 2022, IJRR

Opportunities of Real Human Feedback

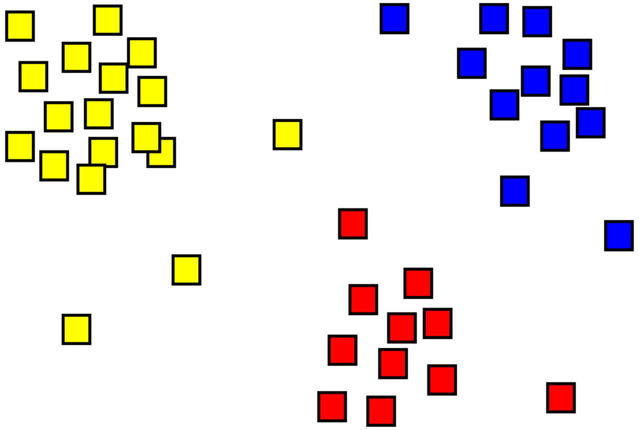
- Optimize the labeling task via:
  - More efficient query selection and presentation[3]
  - Aided evaluation
- Using implicit feedback

[3] Zhang et al., 2022, NeurIPS;

Images: Leonardo.Ai, Wikipedia Commons

Opportunities of Real Human Feedback

- Optimize the labeling task via:
  - More efficient query selection and presentation[3]
  - Aided evaluation
- Using implicit feedback

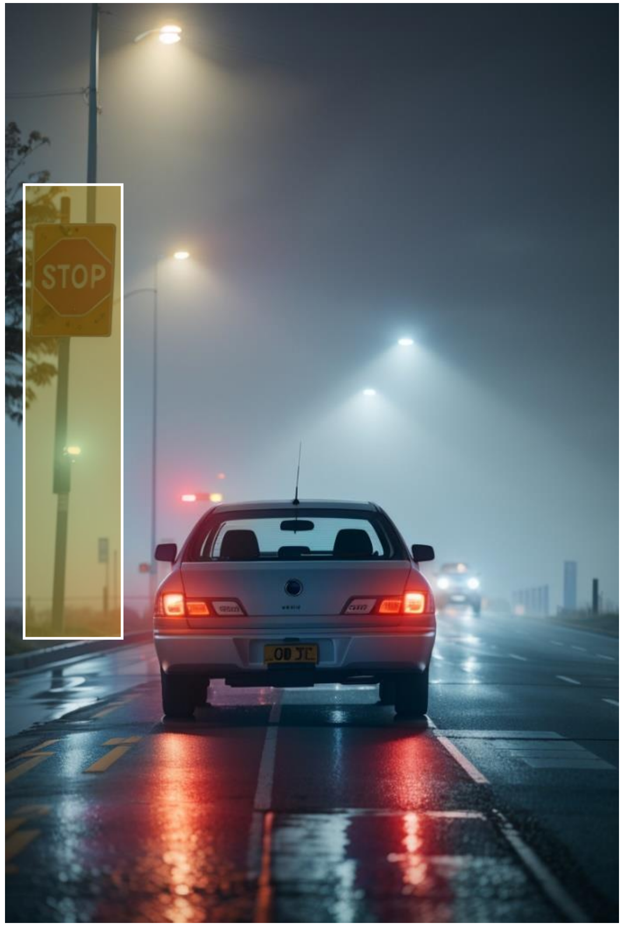


[3] Zhang et al., 2022, NeurIPS;

Images: Leonardo.Ai, Wikipedia Commons

Opportunities of Real Human Feedback

- Optimize the labeling task via:
  - More efficient query selection and presentation[3]
  - Aided evaluation
- Using implicit feedback

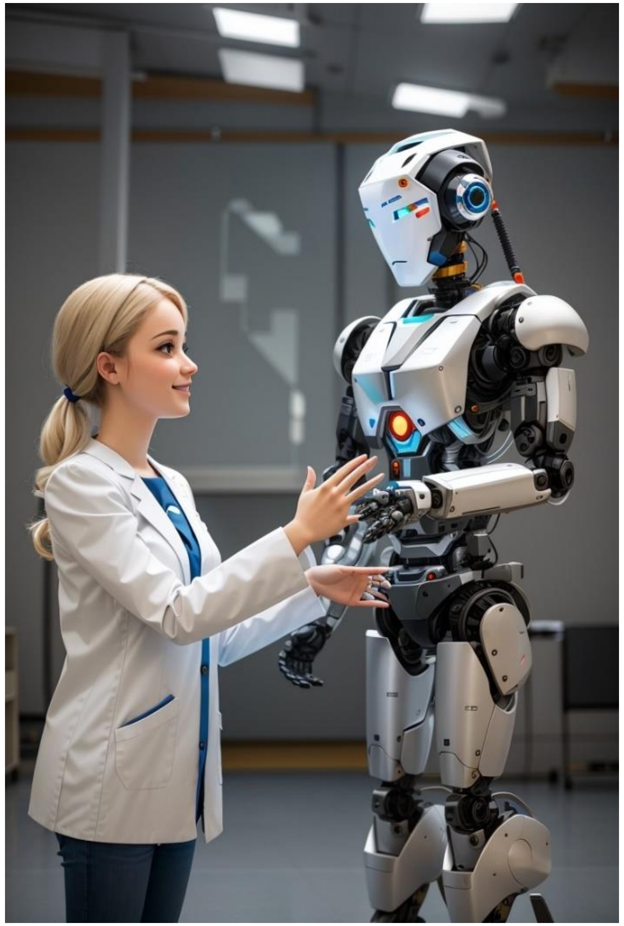


[3] Zhang et al., 2022, NeurIPS;

Images: Leonardo.Ai, Wikipedia Commons

Opportunities of Real Human Feedback

- Optimize the labeling task via:
  - More efficient query selection and presentation[3]
  - Aided evaluation
- Using implicit feedback



[3] Zhang et al., 2022, NeurIPS;

Images: Leonardo.Ai, Wikipedia Commons

Future applications and research ideas

Designing a platform to make collecting HF easier.[3]

Systematically reviewing research on best practices in collecting HF.

→ Facilitate this with platform.

Facilitate collaboration across disciplines to enhance research in RLHF.

[3] Metz et al., 2023, ILHF Workshop ICML

Take-Away

- Synthesized feedback misses crucial aspects of real feedback.
- Real feedback poses challenges, but also provides opportunities.
- It is important to incorporate these aspects into RLHF research.
- We need more research to systematically compare different feedback modes.

Image: Freepik.com

Take-Away

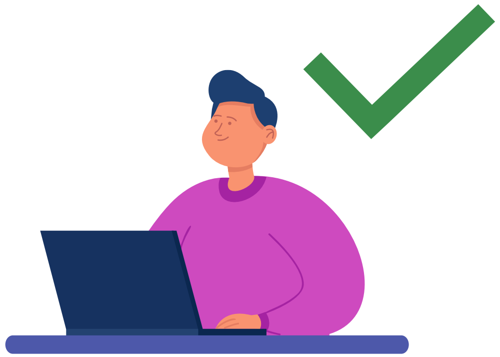
- Synthesized feedback misses crucial aspects of real feedback.
- Real feedback poses challenges, but also provides opportunities.
- It is important to incorporate these aspects into RLHF research.
- We need more research to systematically compare different feedback modes.

More at our poster and online:
timokaufmann.com

Image: Freepik.com

Take-Away

- Synthesized feedback misses crucial aspects of real feedback.
- Real feedback poses challenges, but also provides opportunities.
- It is important to incorporate these aspects into RLHF research.
- We need more research to systematically compare different feedback modes.



More at our poster and online:
timokaufmann.com

Questions?

Image: Freepik.com