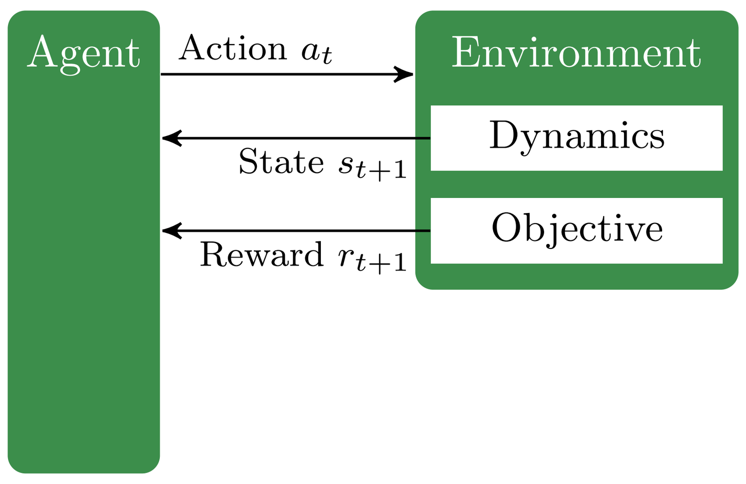
On the Challenges and Practices of Reinforcement Learning from Real Human Feedback

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

LMU Munich
{timo.kaufmann,eyke}@ifi.lmu.de
{sarah.ball,jacob.beck,frauke.kreuter}@stat.uni-muenchen.de

Reinforcement Learning

- **Reinforcement Learning (RL)**: Learning behavior from rewarded interaction with an environment.

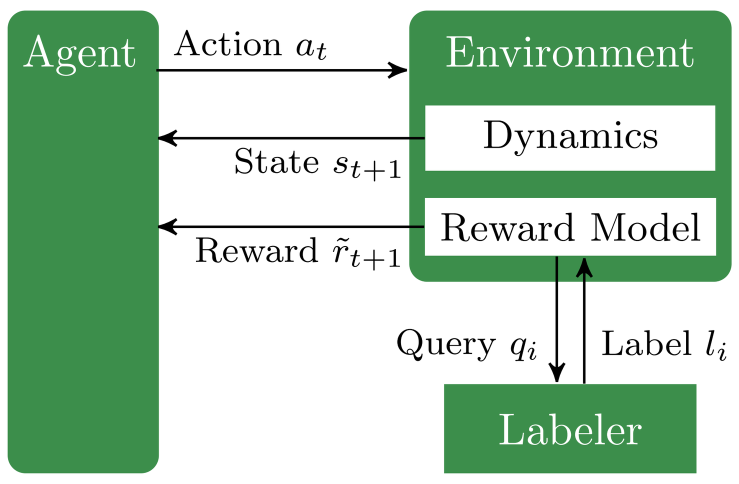


- **Goal:** Find policy π that maximizes

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

From Human Feedback

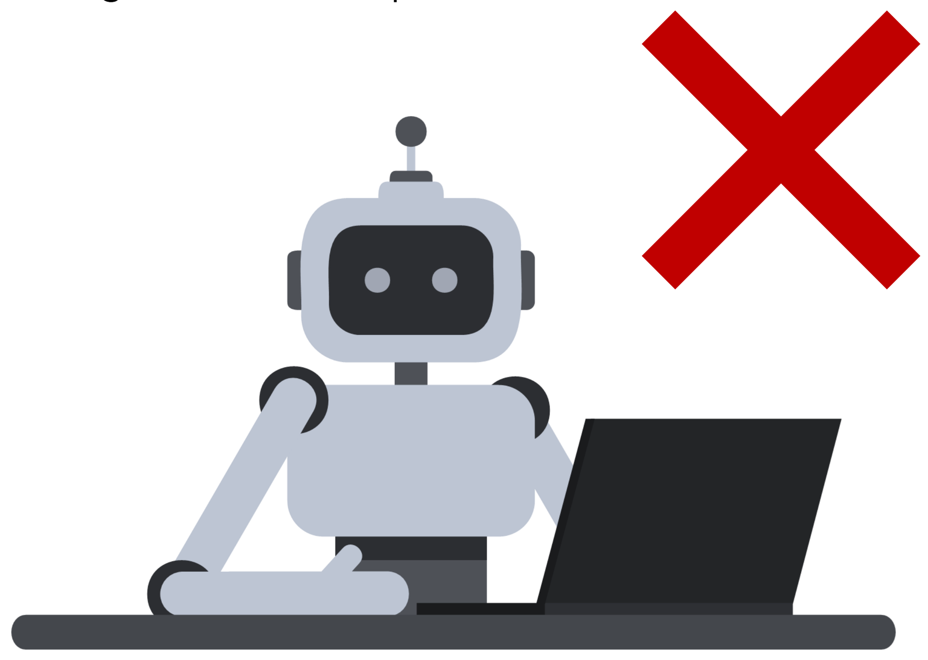
- Defining rewards that induce desired behavior is challenging[1] → **Reinforcement Learning from Human Feedback (RLHF).**



- Many successful applications, e.g., games[1], continuous control[1], instruction fine-tuning[2] (ChatGPT), etc.

Or Synthetic Feedback?

- Real human feedback is inconvenient.
- Researchers often synthesize feedback for evaluation[3].
- Our argument: This is problematic!



Challenges of Real Human Feedback

- **Response biases**, such as acquiescence bias[4], primacy/recency effects[5], satisficing[4] and straightlining[6], may invalidate the human choice model.
- **Motivation** may aggravate or weaken response biases.
- **Fatigue** leads to decreasing label quality over time (*intra-labeler disagreement).*
- **Experience** leads to *increasing* quality over time (intra-labeler disagreement).
- **Misunderstandings** may invalidate feedback and lead to *researcher-labeler disagreement*[2].
- **Expertise** may lead to varying responses from different labelers (*inter-labeler disagreement*).
- **Distractions** may reduce data quality and introduce inconsistencies.
- **Uneven labeling rate** may violate RLHF algorithm assumptions[1].

Opportunities of Real Human Feedback

- **Extensions to comparison queries** such as
  - explanations[7],
  - additional response options[8] and
  - longer interactions,

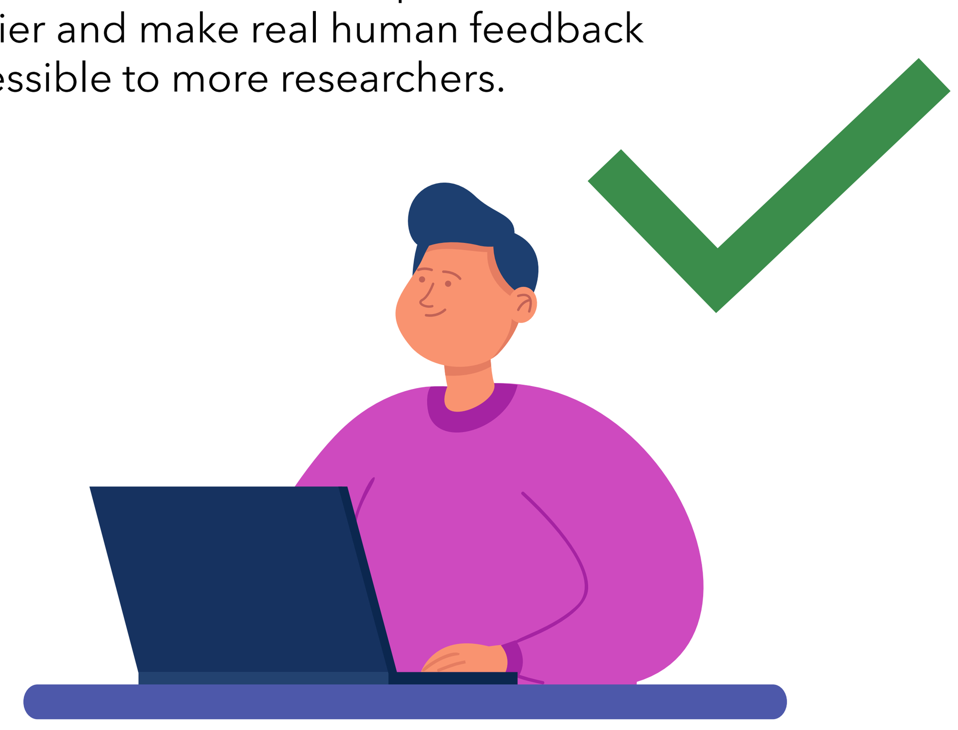
  may provide additional information.
- Evaluating **alternative feedback modalities** may lead to more natural ways of communicating preferences.
- Developing techniques for **aided evaluation** may allow us to leverage the human's strengths[8].
- Optimizing **query presentation** may simplify the feedback task[9].
- Optimizing **query selection** may generate easier questions[10].
- Using **implicit feedback** may provide additional labels for free.
- **Implicit reward shaping** may aid the RL algorithm in learning[1].

User Study Design Decisions

- The **order** of queries is important, especially considering effects such as fatigue and experience.
- Detailed **guidelines** can help to reduce inter-labeler and researcher-labeler disagreements.
- **Incentives** should be well-aligned with the researcher's goals to avoid aggravating response biases.
- **Quality control** can help reduce the impact of response biases and misaligned incentives.
- Careful **participant selection** can supplement quality control and is especially important in crowd-sourcing settings.
- **Interface-driven limitations** such as occlusion of important information can be avoided by careful design of the user-interface.

Take-Away

- Real feedback poses challenges, but also provides opportunities.
- Synthesized feedback misses crucial aspects of real feedback.
- It is important to incorporate these aspects into RLHF research.
- User study design and execution are challenging.
- Future work should attempt to reduce this barrier and make real human feedback accessible to more researchers.



Acknowledgements

This publication was supported by LMU-excellent, funded by the Federal Ministry of Education and Research (BMBF) and the Free State of Bavaria under the Excellence Strategy of the Federal Government and the Länder as well as by the Hightech Agenda Bavaria. This work is also supported by the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research. Jacob Beck was funded by the Munich Center for Machine Learning (MCML).

References

[1] Christiano et al., 2017, Advances in Neural Information Processing Systems
[2] Ouyang et al., 2022, ArXiv Preprint
[3] Lee et al., 2021, Conference on Neural Information Processing Systems
[4] Groves et al., 2009, John Wiley & Sons Publishing
[5] Murphey et al., 2006, Journal of Computer-Mediated Communication
[6] Herzog and Bachman, 1981, The Public Opinion Quarterly
[7] Guan et al., 2021, Advances in Neural Information Processing Systems
[8] Wilde et al., 2022, Proceedings of the Conference on Robot Learning
[9] Zhang et al., 2022, NeurIPS Workshop on Human in the Loop Learning
[10] Bıyık et al., 2022, Proceedings of the Conference on Robot Learning

Read the paper online!

Timo Kaufmann
https://timokaufmann.com

ONE MUNICH Strategy Forum
Next Generation Human-Centered Robotics

Sarah Ball
@sarahba1010

Images: Freepik.com