Timo Kaufmann, PhD Student

timo.kaufmann@ifi.lmu.de @timokauf Timo Kaufmann

https://timokaufmann.com/

Education

2022 –	**Ph.D., LMU Munich** AI+ML group, Prof. Eyke Hüllermeier Anticipated graduation: 10/2026 Thesis focus: *Aligning AI Systems with Human Values through Preference Learning* Research areas: RLHF, preference modeling, human-AI interaction
2019 – 2022	**M.Sc. Computer Science**, University of Paderborn, Germany Grade: 1.0 (with distinction; U.S. GPA 4.0 equivalent) Thesis title: *Curiosity-Driven Semi-Supervised Reinforcement Learning*
2015 – 2019	**B.Sc. Computer Science**, University of Paderborn, Germany Grade: 1.2 (with distinction; approx. U.S. GPA 3.9 equivalent) Thesis title: *Wireless Virtual Network Embedding using Reinforcement Learning*

Research Publications

The following lists my publications, starting with highlighted works that I feel best represent my research.

Highlighted Works

1. **T. Kaufmann**, Y. Metz, D. Keim, and E. Hüllermeier, "ResponseRank: Data-Efficient Reward Modeling through Preference Strength Learning," in *Proc.* ***NeurIPS***, 2025.

   Leveraging implicit comparison strength rankings for more efficient preference learning in RLHF.

2. **T. Kaufmann**, P. Weng, V. Bengs, and E. Hüllermeier, "A Survey of Reinforcement Learning from Human Feedback," ***TMLR***, 2025.

   A comprehensive survey on RLHF, well-received with over 300 citations.

3. A. Findeis, **T. Kaufmann**, E. Hüllermeier, S. Albanie, and R. Mullins, "Inverse Constitutional AI: Compressing Preferences into Principles," in *Proc.* ***ICLR***, 2025.

   Explaining natural language preference datasets.

4. X. Feng, Z. Jiang, **T. Kaufmann**, E. Hüllermeier, P. Weng, and Y. Zhu, "Comparing Comparisons: Informative and Easy Human Feedback with Distinguishability Queries," in *Proc.* ***ICML***, 2025.

   A first step into a direction I am very excited about: Explicit modeling of preference strength.

5. **T. Kaufmann**, S. Ball, J. Beck, F. Kreuter, and E. Hüllermeier, "On the Challenges and Practices of Reinforcement Learning from Real Human Feedback," in *ECML PKDD HLDM Workshop*, 2023.

   Exploring complexities of human feedback, motivating my interest in accurate preference modeling.

All Publications

1. X. Feng, Z. Jiang, **T. Kaufmann**, E. Hüllermeier, P. Weng, and Y. Zhu, "DUO: Diverse, Uncertain, On-Policy Query Generation and Selection for Reinforcement Learning from Human Feedback," in *Proc.* ***AAAI***, 2025. DOI: 10.1609/aaai.v39i16.33824.
2. A. Findeis, **T. Kaufmann**, E. Hüllermeier, and R. Mullins, *Feedback Forensics: A Toolkit to Measure AI Personality*, Under review, arXiv: 2509.26305. http://feedbackforensics.com/, 2025.

3. S. Dutta, **T. Kaufmann**, et al., “Problem Solving Through Human-AI Preference-Based Cooperation,” *Computational Linguistics*, 2025. DOI: `10.1162/coli.a.19`.

4. **T. Kaufmann**, J. Blüml, Q. Delfosse, K. Kersting, and E. Hüllermeier, “OCALM: Object-Centric Assessment with Language Models,” in *RLC 2024 Workshop RLBRew*, 2024.

5. T. Yamagata, T. Oberkofler, **T. Kaufmann**, V. Bengs, E. Hüllermeier, and R. Santos-Rodriguez, “Relatively Rational: Learning Utilities and Rationalities Jointly from Pairwise Preferences,” in *ICML 2024 Workshop on Models of Human Feedback for AI Alignment (MHFAIA)*, 2024.

6. **T. Kaufmann**, V. Bengs, and E. Hüllermeier, “Reinforcement Learning from Human Feedback for Cyber-Physical Systems: On the Potential of Self-Supervised Pretraining,” in *Proceedings of the International Conference on Machine Learning for Cyber-Physical Systems (ML4CPS)*, Springer Nature Switzerland, 2023. DOI: `10.1007/978-3-031-47062-2_2`.

Software & Tools

2025	**Feedback Forensics**: Open-source toolkit for analyzing preference data and AI model personality. Co-developed with A. Findeis (lead developer). GitHub: `rdnfn/feedback-forensics`

Miscellaneous Experience

Academic Service

2023	Organization of an **invited session** on RLHF at the DSSV-ECDA 2023 conference.

Teaching

2025	Supervision of a **software engineering practical** on AI for students with an AI minor.
2024 – 2025	Conception and supervision of a **software engineering practical** on RLHF.
2024	Teaching Assistant for the course **Preference Learning and Ranking**.
ongoing	Supervision of Bachelor’s and Master’s **theses**.

Skills

Languages	Strong reading, writing and speaking competencies in **English** and **German**.
ML & AI Alignment	Preference learning, reinforcement learning from human feedback, preference data interpretation, reward modeling.
Coding	Python, Java, Bash, LaTeX, …
Misc.	Academic research and writing, teaching, small-scale management.

References

PhD Advisor	Prof. Eyke Hüllermeier, LMU Munich	`eyke@lmu.de`
Co-Advisor	Dr. Viktor Bengs, Research Scientist, State of Hesse, Germany	`viktor.bengs@dfki.de`
Coauthor	Prof. Paul Weng, Duke Kunshan University	`paul.weng@dukekunshan.edu.cn`