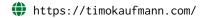
Timo Kaufmann, PhD Student

y @timokauf

in Timo Kaufmann



Education

Ph.D., LMU Munich AI+ML group, Prof. Eyke Hüllermeier 2022 -

Anticipated graduation: 10/2026

Thesis focus: Aligning AI Systems with Human Values through Preference Learning

Research areas: RLHF, preference modeling, human-AI interaction

M.Sc. Computer Science, University of Paderborn, Germany 2019 - 2022

Grade: 1.0 (with distinction; U.S. GPA 4.0 equivalent)

Thesis title: Curiosity-Driven Semi-Supervised Reinforcement Learning

B.Sc. Computer Science, University of Paderborn, Germany 2015 - 2019

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

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.

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.

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.

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.

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

- 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
- 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.



- S. Dutta, **T. Kaufmann**, et al., "Problem Solving Through Human-AI Preference-Based Cooperation," *Computational Linguistics*, 2025. ODI: 10.1162/coli.a.19.
- **T. Kaufmann**, J. Blüml, Q. Delfosse, K. Kersting, and E. Hüllermeier, "OCALM: Object-Centric Assessment with Language Models," in *RLC 2024 Workshop RLBRew*, 2024.
- 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.
- **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. ODI: 10.1007/978-3-031-47062-2_2.

Software & Tools

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

Teaching

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.

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, &TEX, ...

Misc. Academic research and writing, teaching, small-scale management.

References

PhD AdvisorProf. Eyke Hüllermeier, LMU Municheyke@lmu.deCo-AdvisorDr. Viktor Bengs, Research Scientist, State of Hesse, Germanyviktor.bengs@dfki.deCoauthorProf. Paul Weng, Duke Kunshan Universitypaul.weng@dukekunshan.edu.cn