Thomas Kleine Buening

Andreasstrasse 5, 8050 Zurich, Switzerland

⌂ thomasklbg.github.io | 🎓 Google Scholar

Research Experience

ETH Zurich   *Zurich, Switzerland*
ETH AI Center Postdoctoral Fellow '25   *Sept 2025 - current*
- Advised by Andreas Krause and Giorgia Ramponi
- Topics: Reinforcement Learning, LLM Training and Alignment

The Alan Turing Institute   *Oxford, United Kingdom*
Postdoctoral Research Associate   *Mar 2024 - Aug 2025*
- Advised by Marta Kwiatkowska and Lukasz Szpruch
- Topics: Reinforcement Learning

University of Oxford   *Oxford, United Kingdom*
Research Associate, Department of Computer Science   *Mar 2024 - Aug 2025*
- Trinity College Oxford (MCR)
- Topics: Reinforcement Learning

University of British Columbia   *Vancouver, Canada*
Visiting Student in the Probability Group   *Jun 2019 - Dec 2019*
- Hosted by Edwin Perkins
- Topics: Stochastic Calculus
- GPA: 94/100

Education

University of Oslo   *Oslo, Norway*
PhD in Computer Science   *Jan 2021 - Feb 2024*
- Advised by Christos Dimitrakakis
- Topics: Machine Learning, Reinforcement Learning
- Thesis: Learning in the Presence of Cooperative, Adversarial and Strategic Agents

University of Münster   *Münster, Germany*
MSc in Mathematics with Distinction   *Oct 2018 - Sept 2020*
- Topics: Stochastic Processes, Stochastic Calculus
- Thesis: Branching in Random Environment and Linear Fractional Distributions

University of Münster   *Münster, Germany*
BSc in Mathematics   *Oct 2014 - Sept 2017*
- Minor in Computer Science
- Thesis: Optimal Stopping in Discrete Time

Publications and Preprints

Stackelberg Learning from Human Feedback: Preference Optimization as a Sequential Game
  Barna Pasztor, **Thomas Kleine Buening**, Andreas Krause
  preprint, 2025, [pdf]

Strategyproof Reinforcement Learning from Human Feedback
  Thomas Kleine Buening, Jiarui Gan, Debmalya Mandal, Marta Kwiatkowska
  Advances in Neural Information Processing Systems (NeurIPS), 2025, [pdf]

A Unifying Framework for Causal Imitation Learning with Hidden Confounders
  Daqian Shao, **Thomas Kleine Buening**, Marta Kwiatkowska
  SCSL Workshop at ICLR 2025, 2025, [pdf]

A Minimax Approach to Ad Hoc Teamwork
  Villin Victor, **Thomas Kleine Buening**, Christos Dimitrakakis, Haifeng Xu
  International Conference on Autonomous Agents and Multiagent Systems (AAMAS), 2025, [pdf]

Strategic Linear Contextual Bandits
  Thomas Kleine Buening, Aadirupa Saha, Christos Dimitrakakis, Haifeng Xu
  Advances in Neural Information Processing Systems (NeurIPS), 2024, [pdf]

Environment Design for Inverse Reinforcement Learning
  Thomas Kleine Buening*, Villin Victor*, Christos Dimitrakakis
  International Conference on Machine Learning (ICML), Oral Presentation, 2024, [pdf]

Bandits Meet Mechanism Design to Combat Clickbait in Online Recommendation
  Thomas Kleine Buening, Aadirupa Saha, Christos Dimitrakakis, Haifeng Xu
  International Conference on Learning Representations (ICLR), Spotlight Presentation, 2024, [pdf]

ANACONDA: An Improved Dynamic Regret Algorithm for Adaptive Non-Stationary Dueling Bandits
  Thomas Kleine Buening, Aadirupa Saha
  International Conference on Artificial Intelligence and Statistics (AISTATS), 2023, [pdf]

Interactive Inverse Reinforcement Learning for Cooperative Games
  Thomas Kleine Buening, Anne-Marie George, Christos Dimitrakakis
  International Conference on Machine Learning (ICML), 2022, [pdf]
  Best Paper Award at the Cooperative AI Workshop at NeurIPS 2021

On Meritocracy in Optimal Set Selection
  Thomas Kleine Buening, Meirav Segal, Debabrota Basu, Anne-Marie George, Christos Dimitrakakis
  Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO), 2022, [pdf]
  Best Student Paper Award

Teaching Experience

University of Edinburgh   *Edinburgh, UK*
Guest Lecturer   *Feb 2025 - March 2025*
• Exploration vs. Exploitation Trade-Offs in Online Learning

University of Oslo   *Oslo, Norway*
Teaching Assistant   *Aug 2022 - Jul 2023*
• Reinforcement Learning and Decision Making under Uncertainty

University of Münster   *Münster, Germany*
Teaching Assistant   *Oct 2018 - Sept 2020*
• Probability Theory I
• Mathematical Statistics

Qingdao University of Science and Technology   *Qingdao, China*
Tutorial Lecturer at the German-Chinese Campus   *Mar 2018 - Jul 2018*
• Applied Mathematics Refresher for Engineering Students (in German)

Academic Service and Outreach

Workshop   **Models of Human Feedback for AI Alignment (2nd Edition, ICML 2025)**, Co-Organizer
Workshop   **Beyond the Symbols vs. Signals Debate, 2024 (The Royal Society)**, Rapporteur
Workshop   **Models of Human Feedback for AI Alignment (ICML 2024)**, Main Organizer & Moderator
Reviewing ICML 2022-ongoing, NeurIPS 2022-ongoing (Top Reviewer), ICLR 2023-ongoing, AISTATS 2023

Awards

2022   **Best Paper Award with Student Presenter**, EAAMO 2022   *Washington DC*
2021   **Best Paper Award**, Cooperative AI Workshop at NeurIPS 2021   *Virtual*
2019   **DAAD Scholarship**, Research Visit Scholarship   *Vancouver*

Invited Talks

2025   **The Role of Incentives in AI Alignment**, ML Seminar, Czech Academy of Sciences   *Prague*
2024   **Strategic Interactive Decision-Making**, CS Department, University of Chicago   *Chicago*
2024   **Minimax-Bayes Reinforcement Learning**, QAVAS, University of Oxford   *Oxford*

Programming

Python (NumPy, Pandas, scikit-learn, PyTorch), SQL

Languages

German (Native)   **English** (Fluent)   **Chinese (Mandarin)** (Conversational)