LEON ESHUIJS

PhD Student

leoneshuijs [cat - c] gmail [spot, sp=d] com
watermeleon.github.io leoneshuijs
Amsterdam, The Netherlands

EDUCATION

PhD Student

Vrije Universiteit Amsterdam

Oct 2023 - present | Amsterdam, Netherlands

- Research on applying Reinforcement Learning to specialize NLP models
- Recently incorporating Mechanistic Interpretability
- In collaboration with the Universiteit Utrecht

Master Physics - M1

Sorbonne Université

Sep 2022 - Jan 2023 | Paris, France

- Erasmus following the first semester of the Physics Master.

Master Artificial Intelligence

University of Amsterdam

Sep 2020 - Apr 2023 | Amsterdam, Netherlands

- **Electives**: Specialized in advanced Machine Learning, Game Theory, Computer Vision and Reinforcement Learning.
- **Thesis**: "Knowledge Injection through Prompting (KIP): improving image captioning by leveraging knowledge graphs for Transformers". Grade: 8.0/10.0
- **GPA**: 8.3/10.0 - Cum Laude

Bachelor Artificial Intelligence

University of Amsterdam

Sep 2017 - Aug 2020 | Amsterdam, Netherlands

- For the honours program I followed courses in Physics and Big History and Psychopharmacology.
- **Minor:** Electronics for Robotics - TU Delft
- **Thesis:** Reinforcement learning for PID tuning of robot arm.
- **GPA:** 8.0/10.0 - graduated Cum Laude and with Honours

EXPERIENCE

AI Engineer

Saivvy

Sep 2020 – May 2021 | Amsterdam, Netherlands

- Maintained and expanded the code base for the computer vision projects, including object detection.
- Assisted in the coordination and supervision of Master student projects.
- Constructed the control framework for a harvest robot in simulation.

PROJECTS

AI Safety Camp - Unsearch

Jan 2024 - Apr 2024

- As a participant in the AI Safety Camp, I joined the UnSearch team to help understand search in Transformer models
- Investigated path following and world models in Transformer models for mazes via Mechanistic Interpretability technique Path Patching.

SpotMicroAI: RL for gait modulation - Sim2Real gap

Jun 2022 - Feb 2023

- Based on the open-source SpotmicroAI platform, I 3D printed and built my robot dog, Jake, and then trained different Reinforcement Learning (RL) algorithms for the task of gait modulation.
- See website for more information and videos: link

EVENTS

Summer School: Human-aligned AI
Participant July 2024 | Prague, Czech Republic

Hackathon: Deception Detection *(link)*
Participant Jun 2024 | Online - Apart Research

Conference: EAGxUtrecht - Effective Altruism
Participant Nov 2023 | Utrecht, Netherlands

Hackathon: Enhancing Research Productivity *(link)*
Participant - 2nd place Feb 2024 | Kaiserslautern, Germany

Conference: BNAIC *(link)*
Demo and Poster Presentation Nov 2023 | TU Delft

TECHNICAL SKILLS

Experienced with

Python | Pytorch | Keras | Slurm | Conda

Familiar with

C | R | SQL | MATLAB | HTML | CSS | Docker | Azure Machine Learning | GCP

PERSONAL INTERESTS

Memory Athlete | Robotics | Effective Altruism | Inclusive AI

LANGUAGES

Dutch
English
French (B1)

PUBLICATIONS

Leon Eshuijs, Shihan Wang, and Antske Fokkens. "**Short-circuiting Shortcuts: Mechanistic Investigation of Shortcuts in Text Classification**", The SIGNLL Conference on Computational Natural Language Learning, 2025. (link)

Leon Eshuijs, Shihan Wang, and Antske Fokkens. "**Balancing the Scales: Reinforcement Learning for Fair Classification.**" *under review*. (link)

Leon Eshuijs, Archie Chaudhury, Alan McBeth, Ethan Nguyen. "**But what is your honest answer? Aiding LLM-judges with honest alternatives using steering vectors.**", Under Review.

Niklas Höpner, Leon Eshuijs, Dimitrios Alivanistos, Giacomo Zamprogno, Ilaria Tiddi. "**Automatic Evaluation Metrics for Artificially Generated Scientific Research.**" NAACL Workshop on AI and Scientific Discovery, 2025 - Non-Archival. (link)

Leon Eshuijs, Gijs de Jong, and Arnoud Visser. "**Demonstrating reinforcement-learned gaits with two small quadrupeds.**" Proceedings of the 35th Benelux Conference on Artificial Intelligence (BNAIC), 2023. (link)

Gijs de Jong, Leon Eshuijs, and Arnoud Visser. **"Learning to walk with a soft actor-critic approach.**", Proceedings of the 35th Benelux Conference on Artificial Intelligence (BNAIC), 2023. (link)