Alexander Müller

LinkedIn GitHub Google Scholar
alexanderakm.github.io

alexanderklavermuller@gmail.com
Mobile: +31 06 22047760

Profile

MSc Artificial Intelligence student at the University of Groningen with research interests in AI safety, such as mechanistic interpretability and multi-agent systems. Co-directing the AI Safety Initiative Groningen and publishing work on LLM compliance benchmarking [1], mechanistic interpretability [2], activation steering [3], and multi-agent deliberation [4]. Experienced educator ranked 3rd out of ∼50 teaching assistants at the University of Groningen, Faculty of Science and Engineering, in 2025.

Education

- **University of Groningen** Groningen, The Netherlands
  Bachelor of Science – Artificial Intelligence *September 2022 – July 2025*
  - **Relevant Courses**: Algorithms and Data Structures, Object-Oriented Programming, Linear Algebra, Statistics, (Multivariable) Calculus, Neural Networks, Reinforcement Learning, Uncertainty in ML, Analysis, Advanced Logic, Ethics in AI, Cognitive Psychology, Neuroscience
  - **GPA**: 8.9/10
  - **Thesis**: Machine Learning for Peptide Property Prediction: A Robust Nested Cross-Validation Pipeline with Mechanistic Interpretability. Conducted the first case study of mechanistic interpretability in the domain of chemistry. Grade: 9/10. This work led to a follow-up publication [2].

  Master of Science – Artificial Intelligence *September 2025 – July 2027*
  - **Relevant Courses**: Advanced Machine Learning, Deep Learning, Proof Theory and Modal Logic, Logical Aspects of Multi-Agent Systems, Responsible AI, Design of Multi-Agent Systems, Computational Game Theory, Pattern Recognition, Methodology in AI
  - **GPA**: 8.7/10

Publications

[1] I. Lichkovski, **A. Müller**, M. Ibrahim, T. Mhundwa. "EU-Agent-Bench: Measuring Illegal Behavior of LLM Agents Under EU Law.", 2025.

[2] **A. Müller**, J. Cardenas-Cartagena, R. Pollice. "Uncovering Internal Prediction Mechanisms of Transformer-Based Chemical Foundation Models.", 2025.

[3] S. Abreu, J. Postmus, **A. Müller**, J.L. Ferrao, I. Lichkovski, K.F. Michalak, et al. "From Steering Vectors to Conceptors: Compositional Affine Activation Steering for LLMs." 2025.

[4] **A. Müller**, A. Golicins, G. Lesnic. "Collective Deliberation for Safer CBRN Decisions: A Multi-Agent LLM Debate Pipeline.", 2025.

Working Experience

- **Co-Director, AI Safety Initiative Groningen (AISIG)**: Aug. 2025 – Ongoing
  - **Leadership**: Co-directing a student-led AI safety organization of ∼20 team members, ∼300 community members, and ∼1,000 LinkedIn followers. Running education programs, workshops, seminars, and reading groups.
  - **Research**: Facilitating interdisciplinary research published at NeurIPS and ICLR. Participated in projects on LLM compliance [1], activation steering [3], and multi-agent safety [4].
  - **Partnerships**: Collaborating with multiple organizations on AI safety, including Dutch municipality Westerkwartier. Supporting local AI safety initiatives across the Netherlands such as AI Safety Initiative Amsterdam and striving towards a national-level organization.
  - **Outreach**: Represented AISIG at EAGxAmsterdam and aiGrunn conferences with talks on AI safety. Delivering various other talks on AI Safety whenever the opportunity arises.
- **Lead, NeurAlignment Research & Discussion Group**: Nov. 2024 – Nov. 2025
  - **Research Coordination**: Leading a group (BSc–PhD) exploring how neuroscience can inform AI alignment. Currently supervising three research projects.
- **Teaching Assistant**: Sep. 2023 – Feb. 2025
  - **Courses**: Introduction to Artificial Intelligence (2×), Basic Scientific Skills (2×), Cognitive Psychology, Introduction to the Brain. Led tutorials, labs, and assessments across six course instances.
  - **Recognition**: Awarded **3rd place Teaching Assistant of the Year** (out of ∼50 TAs) at the University of Groningen, Faculty of Science and Engineering, in 2025. Highly positive feedback when anonymous feedback forms were filled in.
- **Mentor for First-Year AI Students**: Sep. 2023 – Feb. 2024; Sep. 2024 – Feb. 2025
  - **Student Support**: Mentored ∼20 first-year AI students across two cohorts, guiding them through the Dutch university system and student life.
- **Private Tutor**: Mar. 2019 – Feb. 2024
  - **Personalized Instruction**: Tutored 10+ high school students (some for multiple years) in Mathematics, Chemistry, and Physics over five years, developing tailored lesson plans for the major national exams.

Updated February 19, 2026

Selected Projects

- **The Wisdom of the LLM Crowd**: Can collective LLM deliberation produce moral judgments more aligned with human preferences than individual models? We simulated deliberative groups of five LLMs on 1,000 Moral Machine dilemmas and found that deliberation actually *worsened* moral alignment. Debate-induced extremization collapsed nuanced priors into simplistic heuristics. Project done for the course "Design of Multi-Agent Systems".
- **Intervention Analysis on the Latent Space of VAEs**: How does training data diversity influence uncertainty and reconstruction blur in VAEs? We varied class and sample diversity on MNIST/FashionMNIST, then applied intervention analysis to identify which latent dimensions contribute most to reconstruction error. More diverse data leads to richer, more interpretable latent spaces. Project done for the course "Advanced Machine Learning".
- **Strategic Voting via Dynamic Epistemic Logic**: Modeling strategic voting in repeated elections using dynamic epistemic logic. We model polls as public announcements in S5 epistemic logic and use the Kendall Tau distance as a manipulation heuristic, showing how a strategic voter can successfully manipulate outcomes while inadvertently creating common knowledge. Project done for the course "Logical Aspects of Multi-Agent Systems".
- **Sequent Calculus Proofs for Modal Logic**: Individual assignment on sequent calculus derivations for propositional and modal logic. Covers proofs of standard axioms, cut elimination, and derivations in systems including S4 and S5 modal logic. Project done for the course "Proof Theory and Modal Logic".

Skills

- **Languages**: Dutch (native), English (C2 — Cambridge certificate), German (B2)
- **Programming**: Python, C, Java, R
- **ML & Data Science**: PyTorch, Scikit-learn, Keras, Pandas, NumPy, Matplotlib, Gymnasium
- **Tools**: Git/GitHub, LaTeX, Jupyter
- **Writing**: Academic writing (C2 Cambridge certificate) and writing on two blogs (Personal Substack and AISIG's Substack)

Talks

- **AI Star of the North – Samenwerking Noord (Feb. 2026)**: Featured as "AI-ster van het Noorden" in AI Actueel, the biweekly AI magazine of Samenwerking Noord, discussing the safe application of AI.
- **How to Run an AI Safety Initiative? (Dec. 2025)**: One-hour workshop for 30 attendees at EAGxAmsterdam on founding and scaling a university AI safety initiative.
- **AI Safety at aiGrunn Conference (Nov. 2025)**: 30-minute talk for 50 software developers on risks from advanced AI, focusing on superintelligence.
- **AI Safety at Turn.io (Aug. 2025)**: One-hour talk and Q&A for 30+ staff on risks from LLM-powered chatbots, with actionable mitigation proposals.

Updated February 19, 2026