

Alexander Müller

[in](#) LinkedIn [G](#)itHub [G](#)oogle Scholar
[alexanderkm.com](#)

alexanderklavermuller@gmail.com

Mobile: +31 6 22047760

PROFILE

Director of [Safe AI Netherlands \(SAIN\)](#), the national AI safety field-building organization focused on preparing the Netherlands for the risks of transformative AI, recently funded with a **\$2M two-year grant**. I lead a team of 4 full-time staff (including myself), 10 part-timers, and ~30 volunteers across chapters in Groningen, Utrecht, and Amsterdam, educating hundreds of people, and am expanding SAIN to Eindhoven, Delft, and The Hague. Co-founding and co-directing the AI Safety Hub Amsterdam at The Stack, Amsterdam's new AI hub. Previously did research in multi-agent game theory [1], LLM agent evaluations [2], mechanistic interpretability [3], activation steering [4], and multi-agent deliberation [5]. MSc AI at the University of Groningen (paused to lead SAIN full-time); BSc AI *cum laude*.

WORKING EXPERIENCE

- **Director, Safe AI Netherlands (SAIN):** Apr. 2026 – Ongoing
 - **Funding:** Secured a **\$2M two-year grant** (2026), taking SAIN from a ~€20k bridge budget to a professionally funded national organization.
 - **Team & Operations:** Built SAIN's national structure and now lead a team of 4 full-time staff (including myself), 10 part-timers, and ~30 volunteers. Set up the employment model, salary bands, payroll, and a structured hiring process with standardized work tests.
 - **Programs:** Oversee education programs reaching hundreds of people, and the Project Hub (6 supervisors, ~15 researchers) whose work has appeared at NeurIPS and ICLR workshops. Advised Dutch municipality Westerkwartier on an AI governance framework.
 - **Expansion:** Leading SAIN's growth from Groningen, Utrecht, and Amsterdam to new chapters in Eindhoven, Delft, and The Hague.
- **Co-Founder & Co-Director, AI Safety Hub Amsterdam:** 2026 – Ongoing
 - **Funding:** Co-founding a dedicated AI safety hub at The Stack, Amsterdam's new AI hub: a shared home for AI safety researchers, professionals, and organizations in the Netherlands.
- **Co-Director, AI Safety Initiative Groningen (AISIG):** Aug. 2025 – Apr. 2026
 - **Leadership:** Co-directed a student-led AI safety organization of ~20 team members, ~300 community members, and ~1,000 LinkedIn followers. Ran education programs, workshops, seminars, and reading groups. Led AISIG's transformation into SAIN, becoming its founding Groningen chapter in April 2026.
 - **Research:** Facilitated interdisciplinary research published at NeurIPS and ICLR workshops. Participated in projects on LLM compliance [2], activation steering [4], and multi-agent safety [5].
 - **Partnerships:** Collaborated with multiple organizations on AI safety, including Dutch municipality Westerkwartier. Supported local AI safety initiatives across the Netherlands such as AI Safety Initiative Amsterdam.
 - **Outreach:** Represented AISIG at EAGxAmsterdam and aiGrunn with talks on AI safety.
- **Lead, NeurAlignment Research & Discussion Group:** Nov. 2024 – Nov. 2025
 - **Research Coordination:** Led a group (BSc–PhD) exploring how neuroscience can inform AI alignment and supervised three research projects.
- **Teaching Assistant, University of Groningen:** 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), Faculty of Science and Engineering, 2025.
- **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 in Mathematics, Chemistry, and Physics over five years, preparing them for the national exams.

PUBLICATIONS

- [1] A. Todorov, J. Napel, **A. Müller**. "[Parametric Open Source Games](#)." arXiv preprint, 2026.
- [2] I. Lichkovski, **A. Müller**, M. Ibrahim, T. Mhundwa. "[EU-Agent-Bench: Measuring Illegal Behavior of LLM Agents Under EU Law](#)." Workshop paper, 2025.
- [3] **A. Müller**, J. Cardenas-Cartagena, R. Pollice. "[Uncovering Internal Prediction Mechanisms of Transformer-Based Chemical Foundation Models](#)." ChemRxiv, 2025.
- [4] 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](#)." Workshop paper, 2025.
- [5] **A. Müller**, A. Golicins, G. Lesnic. "[Collective Deliberation for Safer CBRN Decisions: A Multi-Agent LLM Debate Pipeline](#)." Apart Research, 2025.

EDUCATION

- **University of Groningen** Groningen, The Netherlands
 - *Master of Science – Artificial Intelligence (paused to lead SAIN)* *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:** 8.8/10
 - *Bachelor of Science – Artificial Intelligence, Cum Laude* *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 (*cum laude*)
 - **Thesis:** Machine Learning for Peptide Property Prediction: A Robust Nested Cross-Validation Pipeline with Mechanistic Interpretability. The first case study of mechanistic interpretability in chemistry. Grade: 9/10. Led to a follow-up publication [3].

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 *worsened* moral alignment: debate-induced extremization collapsed nuanced priors into simplistic heuristics.
- **Strategic Voting via Dynamic Epistemic Logic:** Modeling polls in repeated elections as public announcements in S5 epistemic logic, showing how a strategic voter can manipulate outcomes while inadvertently creating common knowledge.

SKILLS

- **Leadership & Operations:** Fundraising and grant writing, hiring and team management, organizational design, program design, partnerships
- **Research Experience:** Multi-agent game theory, LLM agent evaluations and benchmarks, mechanistic interpretability, activation steering
- **Languages:** Dutch (native), English (C2 — Cambridge certificate), German (B2)
- **Programming & ML:** Python, C, Java, R; PyTorch, Scikit-learn, Pandas, NumPy, Gymnasium; LLM APIs and agentic tooling
- **Tools:** Git/GitHub, $\LaTeX$, Jupyter
- **Writing:** Writing on two blogs ([Personal Substack](#) and [SAIN'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.