

Douw Marx

AI safety researcher

 Leuven, BE
  douwmarx@gmail.com
 0470 51 64 51
  DouwMarx
  [douw-marx-913503163](https://www.linkedin.com/in/douw-marx-913503163)

 Google Scholar
  Equistamp Substack

Summary

AI safety researcher working on agentic evaluations, risk estimation, and alignment. Contributing to the [European Commission Technical Assistance for AI Safety](#), [Redwood Research](#), [AI Futures Project](#), and [METR](#) via [Equistamp](#). PhD from [KU Leuven](#), Belgium, applying machine learning and signal processing to fault detection in rotating machines.

Experience

Equistamp, Research Engineer

- Contributing to [EU AI Office](#) evaluations on CBRN, loss of control, and harmful manipulation
- Contributing to [Redwood Research's LinuxArena](#), a control setting for highly privileged AI agents
- Contributor to [Redwood Research's ASMR-Bench](#), a benchmark for detecting AI sabotage in ML research
- Merged upstream contributions to open-source eval infrastructure: [METR's Hawk](#) and [Redwood Research's Control Tower](#)
- Baselineing for [Andon Labs](#), Redwood Research, and [METR](#)
- Equistamp communications on [Substack](#), including evaluation selection approaches based on risk modelling and value of information
- Scenario evaluations for the [AI Futures Project](#)
- Credited for feedback on [AI 2040: Plan A](#), [AI Futures Project](#)
- Grant writing and associated automations

Leuven, Belgium
Aug 2025 – present
1 year 2 months

KU Leuven, Marie Skłodowska-Curie PhD Fellow

Research on fault detection of rotating machines using unsupervised learning and differentiable signal processing methods.

- [Published 5 peer-reviewed conference papers](#)
- Proposed and led 4 Master's research projects
- Research secondments at INSA Lyon, France and SafranTech, Paris, France
- Runner-up in PHM 2023 Conference Data Challenge, Salt Lake City, Utah, USA

Leuven, Belgium
Sept 2021 – Sept 2025
4 years 1 month

Wolfram MathCore, SystemModeler Intern

[Develop Virtual labs using Wolfram System Modeler and Mathematica.](#)

Linköping, Sweden
May 2021 – July 2021
3 months

XRAM Technologies, Data Analyst and Mechanical Engineer

Build data-driven measurement models for electrode length prediction in electric-arc furnaces.

- Furnace electrode length prediction
- Chemical process data analysis
- Bayesian structural models and Gaussian process regression

South Africa
Jan 2021 – Sept 2021
9 months

Wolfram Summer School, Participant

[Towards universal robotics](#)

- Cellular Automata
- Inverse Kinematics

USA
June 2020 – July 2020
1 month

Publications

Scientific articles	2025
Constitutional Sensitivities of Preference Models	2025
Apart Research Hackathons: AI monitoring at low audit budgets, and Automated risk modelling for evaluation value-of-information estimation	2026
Blog posts on Equistamp's Substack: Value of information and the seven deadly evaluation sins, and No more snapshots: forecast AI risk by backtesting	Sept 2026
Workshop: Agentic AI safety	2026
Talk: Does your LLM care about the same things you do? (75th Data Science Leuven Meetup, slides)	Feb 2026
Patent for a novel parallel kinematic planar mechanism: WO2020208551A1	Feb 2022

Education

PhD KU Leuven, Fault detection in rotating machinery Marie Skłodowska-Curie PhD fellow at KU Leuven. - Thesis: Fault Detection in Rotating Machinery Using Domain Knowledge and Reference Data - Regularised unsupervised deep learning models - Differentiable signal processing methods	Sept 2019 – Oct 2025
M.Eng. University of Pretoria , Honours and Master's in Mechanical Engineering - Machine learning applied to vibration monitoring - Numerical methods and optimisation - Thesis: Towards a hybrid approach for diagnostics and prognostics of planetary gearboxes	Jan 2019 – Dec 2020
B.Eng. University of Pretoria , Mechanical Engineering (With distinction)	Jan 2015 – Dec 2018

Skills and Tools

Evaluation infrastructure

METR's Hawk, Control Tower

Programming/Mathematical Modelling

Python (PyTorch, Pandas, SciPy, Dash, LangChain, Scikit-learn), Matlab, C++, Mathematica

CLI Tools

HPC with Slurm, Linux, LaTeX, Git, Vim, Emacs

Languages

Fluent in English and Afrikaans, understands Dutch.

Certificates

AI Alignment Course: [Bluedot](#) technical alignment course with project: [Constitutional sensitivities of preference models](#).

EA Introductory Program: [Effective Altruism Introductory Program](#)

Interests

Effective altruism, music, woodworking, hiking