Aengus Lynch, Ph.D

aenguslynch@gmail.com — @aengus_lynch1 — Google Scholar — GitHub — aenguslynch.com

Generalist operator and AI safety researcher. First author of Anthropic's blackmail study and its 2026 follow-up.

EDUCATION

PhD in Artificial Intelligence 2021–2025
University College London
Advisor: Prof. Ricardo Silva

MSci Mathematics, First Class Honours 2017–2021
University College London
Master's Project: Computations and Analysis on Fluid Flow Through a Flexible Channel

EXPERIENCE

Researcher June 2026–present
Theorem, San Francisco
Building formally verified software environments as RL training data for frontier AI models.

- Designed and shipped a paid suite of RL environments, delivered on schedule to a frontier lab customer.
- Wrote the technical sales document that scoped the offering, covering difficulty calibration and reward-hack mitigations, and proposed follow-on contracts.
- Up-skilled in formal verification (Rocq, Lean, machine-checked binary equivalence proofs) from a standing start.

Anthropic Fellow Jan 2026–April 2026
First author of "Agentic Misalignment in Summer 2026" (Anthropic Alignment Science, July 2026), documenting four failure modes in frontier models: covert sabotage, assisting fraud, motivated mislabeling, and coaching human proxies to whistleblow. The motivated mislabeling experiments showed LLM judges knowingly flipping labels based on downstream training consequences, a threat to AI-supervised training pipelines.

Founder and CEO February 2025–December 2025
Watertight AI
Founded a third-party AI auditing company selling alignment evaluations and monitoring to frontier labs.

- Closed and executed an eight-week technical research contract with Anthropic to develop evaluation datasets.
- Wound the company down and returned capital after concluding I could not hire the research team the mission required and product-market fit had not yet arrived.

Anthropic contractor August 2024–April 2025
Led the research behind "Agentic Misalignment: How LLMs Could Be Insider Threats," stress-testing 16 frontier models for insider-threat behaviors such as blackmail and corporate espionage. The scenarios are now a standard benchmark, adopted into UK AISI's Inspect evals and cited across subsequent scheming and alignment research, with media coverage from BBC, Fortune, and 15+ major outlets. Joint first author on Anthropic jailbreaking research (Best-of-N

Jailbreaking).

MATS Scholar Jan 2024–August 2024
LLM unlearning and adversarial robustness

REMIX program, Redwood Research Jan 2023
Mechanistic interpretability research

Rates Trading Summer Analyst Jul 2020–Aug 2020
JP Morgan, London
UK rates trading desk for index-linked gilts

PUBLICATIONS

Thesis: The Persistent Vulnerability of Aligned AI Systems
Aengus Lynch
PhD thesis, University College London, 2025. Supervised by Ricardo Silva. Examined by Florian Tramèr and Ilija Bogunovic.
arXiv:2604.00324

2026: Agentic Misalignment in Summer 2026
Aengus Lynch, John Hughes, Alex Serrano, Robert Kirk, Samuel R. Bowman
Anthropic Alignment Science Blog, July 2026
alignment.anthropic.com/2026/agentic-misalignment-summer-2026

2025: Agentic Misalignment: How LLMs Could be Insider Threats
Aengus Lynch, Benjamin Wright, Caleb Larson, Stuart J. Ritchie, Sören Mindermann, Evan Hubinger, Ethan Perez, Kevin K. Troy
arXiv:2510.05179

2024: Best-of-N Jailbreaking
John Hughes*, Sara Price*, **Aengus Lynch***, Rylan Schaeffer, Fazl Barez, Sanmi Koyejo, Henry Sleight, Erik Jones, Ethan Perez, Mrinank Sharma
arXiv:2412.03556

Latent Adversarial Training Improves Robustness to Persistent Harmful Behaviors in LLMs
Abhay Sheshadri*, Aidan Ewart*, Phillip Guo*, **Aengus Lynch***, Cindy Wu*, Vivek Hebbar*, Henry Sleight, Asa Cooper Stickland, Ethan Perez, Dylan Hadfield-Menell, Stephen Casper
arXiv:2407.15549

Analyzing the generalization and reliability of steering vectors
Daniel Tan, David Chanin, **Aengus Lynch**, Adrià Garriga-Alonso, Brooks Paige, Dimitrios Kanoulas, Robert Kirk
NeurIPS 2024
arXiv:2407.12404

Eight methods to evaluate robust unlearning in LLMs
Aengus Lynch*, Phillip Guo*, Aidan Ewart*, Stephen Casper, Dylan Hadfield-Menell
arXiv:2402.16835

2023: Towards automated circuit discovery for mechanistic interpretability
Arthur Conmy, Augustine N. Mavor-Parker, **Aengus Lynch**, Stefan Heimersheim, Adrià Garriga-Alonso

NeurIPS 2023 (Spotlight)
arXiv:2304.14997

Spawrious: A benchmark for fine control of spurious correlation biases
Aengus Lynch*, Gbètondji J-S Dovonon*, Jean Kaddour*, Ricardo Silva
arXiv:2303.05470

2022: Causal machine learning: A survey and open problems
Jean Kaddour*, **Aengus Lynch***, Qi Liu, Matt J. Kusner, Ricardo Silva
arXiv:2206.15475