Last updated in May 2025

Alexander F. Spies

London, UK afspies@imperial.ac.uk afspies.com afspies afspies

Summary

PhD candidate specialising in **mechanistic interpretability** and representation learning for safer advanced AI systems. Published work on object-centric reasoning, causal world models and transformer analysis. Now seeking to work on technical alignment in applied settings.

Education

Imperial College London, Computer Science (Artificial Intelligence) *London, UK*
Thesis — *Interpretable Representations in Artificial Neural Networks* *Oct 2020 – present*

- Improving representations in Object-Centric Learning and reasoning.
- Mechanistic analysis of vision & language transformers

Imperial College London, Computing (AI & ML) *London, UK*

- Thesis - *Unsupervised World Models in the Animal-AI Environment* *Sept 2019 – Sept 2020*
- Independent project - *Neurosymbolic Learning & Neurally Weighted dSILP*

University of California, Berkeley, Major: Physics *Berkeley, CA, USA*

- Completed graduate-level courses as an undergrad, alongside research *Aug 2017 – May 2018*

University of Manchester, Physics (Theoretical) *Manchester, UK*

- Thesis — *AI for the Automated Diagnosis of Atrial Fibrillation* *Sept 2015 – June 2019*

Experience

Research Engineer Intern *London, UK*
Epic Games *Jan 2025 – present*

Research-engineer intern focused on large-scale fine-tuning of LLMs on low-resource languages.

- Implemented finetuning pipeline for local as well as cloud-based training (UnSloth, SageMaker, etc.).
- Deployed evaluation suite with W&B sweeps, vLLM serving and multiple LLM APIs.

Research Team Lead *Remote*
UnSearch (AI Safety Camp) *Mar 2023 – Oct 2024*

Led independent research groups on language model behaviour and interpretability.

- Developed research agenda on mechanistic interpretability for understanding maze-solving LLMs.
- Trained transformers models and Sparse Autoencoders and developed interpretability pipelines.
- Managed 9 researchers across 2 projects resulting in 2 workshop papers and a best-poster award.

JSPS Doctoral Fellow *Tokyo, Japan*
National Institute of Informatics *Aug 2023 – June 2024*

Mechanistic analysis of Transformers trained on maze-solving tasks.

Undergraduate Researcher *Berkeley, CA, USA*
Lawrence Berkeley National Laboratory *Feb 2018 – July 2018*

Investigated non-local thresholds in pixel detectors; co-authored JINST publication.

Research Intern *Hamburg, Germany*
German Electron Synchrotron (DESY) *July 2018 – Sept 2018*

Exclusion analysis of Higgs decay channels in MSSM.

Selected Publications

Transformers Use Causal World Models in Maze-Solving Tasks *Oct 2024*
A.F. Spies, W. Edwards, M.I. Ivanitskiy, et al.
arxiv.org/abs/2410.00000 (World Models Workshop (ICLR 2025))

Structured World Representations in Maze-Solving Transformers *Dec 2023*
M.I. Ivanitskiy, ***A.F. Spies***, T. Räuker, et al.
arxiv.org/abs/2312.00000 (Unifying Representations in Neural Models Workshop (NeurIPS 2023))

Sparse Relational Reasoning with Object-Centric Representations *July 2022*
A.F. Spies, A. Russo, M. Shanahan
arxiv.org/abs/2207.12345 (Dynamic Neural Networks Workshop (ICML 2022) — ***spotlight***)

Awards and grants

Long-Term Future Fund Grant — Safe AI Research	*July 2024*
FAR Labs Residency	*June 2024*
Best Poster — Technical AI Safety Conference	*Apr 2024*
JSPS Postdoctoral Fellowship	*May 2023*
Google Cloud Research Grant	*Aug 2022*
Full PhD Scholarship (UKRI)	*Sept 2020*

Leadership & service

Technical Research Advisor *London, UK*
Pivotal Fellowship *Jan 2025 – Apr 2025*
Provided technical guidance on AI Safety Research to 8+ Research Fellows

Research Proposal Reviewer *June 2024 – July 2024*
ML Alignment & Theory Scholars
Evaluated research proposals for alignment-focused projects

Journals & Top ML conferences *2022 – present*
Reviewer
NeurIPS, ICLR, ICML, AAAI, UAI, Artificial Intelligence (Journal)

Teaching Assistant *Sept 2021 – Feb 2025*
Imperial College London & Manchester

- Led technical coursework for Deep Learning, ML Math, Data Structures & Algorithms, and Python
- Engineered GPU-backed autograding pipeline for 120+ students using Otter Grader and Paperspace

Co-founder — ICARL Seminar Series *London, UK*
Imperial College London *Jan 2021 – present*
Organized talks and receptions with field experts in Reinforcement Learning and AI more broadly

Skills

Frameworks & MLOps: TransformerLens, HF Transformers, PyTorch, Jax, Weights & Biases, Pandas

Research Interests: Mechanistic Interpretability, Causal World Models, AI Safety, Representation Learning

Programming: Python, C++, Java, Git, HTML, CSS, JavaScript

Languages: English (native), German (native), Japanese (beginner)