Adrià Garriga-Alonso

Web page: agarri.ga   adria.garriga@gmail.com
in LinkedIn: adrigarriga   Google scholar page   GitHub: `rhaps0dy`

Profile

AI researcher and engineer, focused on alignment and mechanistic interpretability (2022 – present), previously Bayesian ML (2017 – 2021). Senior author on field-defining interpretability work Automated Circuit Discovery ($\sim$370 citations), which sparked many follow-ups.

Combines deep research leadership with hands-on ML engineering: led teams of up to 6 researchers while building and maintaining critical infrastructure (8-80 GPU cluster, multi-node training) that accelerated research velocity at FAR AI. Apollo Research adopted the infrastructure as a reference design, and I consulted for Goodfire AI's infrastructure.

Seeking to advance AI alignment techniques by applying them to building frontier models, with industry-grade data and compute.

Volunteer research mentorship

2024 – present **Program Mentor** *ML Alignment and Theory Scholars*

Advised 10 scholars on mechanistic interpretability and RL. Results: 3 NeurIPS papers, 5 workshop papers, 2 under review. Mentees progressed to positions at Anthropic, METR, and Mistral.

2025 – present **Fellowship Mentor** *Cambridge Boston Alignment Initiative*

Advised 2 junior collaborators on 1 project: investigate whether LLM chain of thought answers are causal or rationalizations, using activation probes. Under review.

Professional Experience

2023 – present **Research Scientist** *FAR AI*

Led two large projects on interpretability, directly managing 3 people and collaborating with 11.

1. Automated Circuit Discovery: FAR AI's most-cited work ($\sim$370 citations), which kicked off a new subfield in algorithmic circuit discovery.
2. Learned Planners. To study goal-directed behavior, we trained a model to play Sokoban using RL. Showed that the model learned to plan and internally represents future actions (ICLR Oral), and finally reverse-engineered the planning algorithm it learned.

Built and currently maintain FAR AI's GPU/LLM infrastructure (8–80 GPUs; >400B parameters), reducing costs by $\sim$50%, enabling $\geq$20 researchers and collaborators to run large-scale experiments. Established core research project structure, engineering tools, and workflows now standard across the organization.

2022 – 2023 **Member of Technical Staff** *Redwood Research*

Led correctness testing for an optimizing compiler and algebra system used by $\sim$40 interns and 5 full-time researchers. Built fuzzing/property-based testing suite that uncovered critical bugs, ensuring reliability for experiments underpinning the Causal Scrubbing paper (co-author), and others. Mentored 8 interns across 4 projects, including Indirect Object Identification.

2021 **Summer research fellow** (open-source game theory) *Center on Long-Term Risk*

What are the results of games when agents can read each other's source code? Proved that Nash-like equilibria are reached by probabilistic agents that attempt to prove each other's behavior, and that such behavior is described by modal logic.

2019 **Research Intern** *Microsoft Research Cambridge*
With Dr. Sebastian Tschiatschek. Introduced a theoretically motivated algorithm to optimally choose and learn from partial observations of the teacher, in inverse reinforcement learning.

2015 **Research Assistant** *Music Technology Group, UPF*
With Prof. Rafael Ramírez. Developed web app to help music students practice. It listens to the student's playing or singing and visually compares the pitches and durations to an expert's.

2014 **CTO, cofounder** *MonkingMe.com*
Music streaming startup. Designed and implemented web application backend and cloud server infrastructure for the streaming service. Coordinated development of smartphone application.

2013 – 2014 **R&D Intern** *Big Arm Applications*
Developed a communication protocol between a web app and an industrial robot arm. The arm interacted with users via web and mobile, using a webcam, and played the metallophone.

Education

2017 – 2021 **PhD in Machine Learning** *Engineering Department, University of Cambridge, UK*
Supervisor: Prof. Carl E. Rasmussen.
Thesis: "Priors in finite and infinite Bayesian convolutional neural networks".
First to show that, at initialization, an infinitely wide non-MLP architecture converges to a Gaussian process. Proposed exact stochastic-gradient MCMC algorithm and improved priors for Bayesian neural networks.

2016 – 2017 **MSc Computer Science** (distinction) *University of Oxford, UK*
Thesis supervisor: Prof. Mihaela van der Schaar.
Thesis: "Probability density imputation of missing data with Gaussian mixture models".

2012 – 2016 **BSc Computer Science** (1st in class of 67) *Pompeu Fabra University, Spain*
Grade: 9.02/10, graduating class average 6.90/10. Thesis supervisor: Prof. Anders Jonsson.
Thesis: "Solving Montezuma's Revenge with planning and reinforcement learning".

Selected Publications

A. Conmy, A. Mavor-Parker, A. Lynch, S. Heimersheim, **A. Garriga-Alonso**. "Towards Automated Circuit Discovery for Mechanistic Interpretability". NeurIPS (Spotlight), 2023. arXiv:2304.14997. ∼370 citations.

A. Garriga-Alonso, L. Aitchison, C.E. Rasmussen. "Deep Convolutional Networks as Shallow Gaussian Processes". ICLR, 2019. arXiv:1810.05148. ∼330 citations.

L. Sharkey, B. Chughtai, […], **A. Garriga-Alonso**, *et al.* "Open Problems in Mechanistic Interpretability". arXiv, 2025. arXiv:2501.16496. ∼60 citations.

A. Srivastava, A. Rastogi, A. Rao, […], **A. Garriga-Alonso**, *et al.* "Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models". Transactions on Machine Learning Research, 2023. arXiv:2206.04615. ∼2,000 citations.

L. Chan, **A. Garriga-Alonso**, N. Goldowsky-Dill, R. Greenblatt, *et al.* "Causal Scrubbing: A Method for Rigorously Testing Interpretability Hypotheses". Alignment Forum, 2022. ∼90 citations.

C. Shi, N. Beltran Velez, A. Nazaret, C. Zheng, **A. Garriga-Alonso**, A. Jesson, *et al.* "Hypothesis Testing the Circuit Hypothesis in LLMs". NeurIPS, 2024. arXiv:2410.13032.

T. Bush, S. Chung, U. Anwar, **A. Garriga-Alonso**, D. Krueger. “Interpreting Emergent Planning in Model-Free Reinforcement Learning”. ICLR (Oral), 2025. arXiv:2504.01871.

Leadership

2019 **Co-organizer,** ICLR 2019 workshop: “Safe Machine Learning”

2017 – 2019 Founded and led **Engineering AI Safety reading group** *University of Cambridge*
Objective: introduce ML students to beneficial AI techniques. 7–50 attendees per session.

Service to the Scientific Community

Reviewer NeurIPS 2019 (**top 5%**), 2020, 2025. ICLR 2020, 2021, 2026. ICML 2020, 2021, 2023, 2025. JMLR. Workshops: ICML 2024 NextGen AI safety, Mechanistic Interpretability (ICML 2024, NeurIPS 2025).
Mentor New in ML workshop, 2019.

Selected Awards & Fellowships

2017 **Malmo Collaborative AI Challenge: 1st & 3rd places, diff. categories.** *Microsoft Research*
Won $20,000 in Azure credits and paid attendance to the AI Summer School.

2016 **María de Maeztu Award** for Reproducibility in Software. *Pompeu Fabra University*
Best computer science Bachelor’s thesis in Spain meeting scientific reproducibility criteria.

2016 – 2017 **Postgraduate fellowship** (6.6% acceptance rate). *“la Caixa” Foundation*
Full tuition and stipend. Awarded Spain-wide, on academic merit and positive impact of project.

Additional Activities

2014 – 2015 **Competitive programming team member** *Pompeu Fabra University*
Conducted unofficial training sessions. Team set record in Pompeu Fabra University for number of problems solved, placing 24/52 in SWERC 2015.

Skills

Extensive experience with both PyTorch (most recently: interpretability projects) and Jax (RL for Sokoban). My freshest systems language is C++, but I can also write Rust.