# Adrià Garriga-Alonso

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

## Experience

2024–present **Mentor, interpretability** *MATS program*
Direct research projects from scholars transitioning their career into mechanistic interpretability and alignment research. Output: 4 ICML workshop papers and 3 research blog posts from my Winter 2024 cohort, 1 blog post from my Summer 2024 cohort of 1.

2023–present **Research Scientist** *FAR AI*
Conceptualize and lead research to ensure that AI is trustworthy and beneficial to society. Main research project: measuring whether AI systems explicitly conceptualize their goals and what those are. Mentor more junior staff in ML engineering and research.

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

- Established property-based testing and fuzzing to ensure correctness of an interpretability algebra system and tensor compiler
- Co-wrote the Causal Scrubbing paper, whose experiments rely on said compiler.
- Interpretability project: characterized a circuit in GPT2-small that processes skips in monotonic number sequences.

2021 **Summer research fellow** *Center on Long-Term Risk*
Advancing open-source game theory by restricting programs to a tractable yet expressive subset. The result should be a library to model cooperation between AIs with different goals.

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 – 2015 **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.

## Education

2017 – 2021 **PhD Engineering** *Machine Learning Group, University of Cambridge, UK*
Supervisor: Prof. Carl E. Rasmussen.
Thesis: "Priors in finite and infinite Bayesian convolutional neural networks".
Found a theoretical connection between deep Bayesian CNN priors and Gaussian processes (GPs), which has been used for NN generalization bounds and neural architecture search. Building on this, I improved priors and inference for Bayesian CNNs and image GPs.

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 of my class of 67) *Pompeu Fabra University, Spain*
Grade: 9.02/10, class grade average 6.90/10. Thesis supervisor: Prof. Anders Jonsson.
Thesis: "Solving Montezuma's Revenge with planning and reinforcement learning".

## Selected Publications

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

Rohan Gupta*, Iván Arcuschin*, Thomas Kwa, **Adrià Garriga-Alonso**. "InterpBench: Semi-Synthetic Transformers for Evaluating Mechanistic Interpretability Techniques". Mechanistic intepretability workshop, ICML 2024.

**Adrià Garriga-Alonso**, Mohammad Taufeeque, Adam Gleave. "Planning behavior in a recurrent neural network that plays Sokoban". Mechanistic intepretability workshop, ICML 2024.

## Leadership

2019 **Coorganiser,** ICLR 2019 workshop: "Safe Machine Learning"
Other organisers include Dr. Silvia Chiappa and Dr. Victoria Krakovna from DeepMind, and Dr. Adrian Weller, head of Safe and Ethical AI at The Alan Turing Institute.

2017 – 2019 Started and ran **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. ICLR 2020, 2021. ICML 2020, 2021, 2023. JMLR. Workshops: ICML 2024 NextGen AI safety, Mechanistic Interpretability.
**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 Bachelor's thesis in information technologies 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 Universitat Pompeu Fabra for number of problems solved, placing 24/52 in SWERC 2015.