Roberto-Rafael Maura-Rivero, PhD

LinkedIn · X · Scholar · Website

CONTACT

Email: rrmaura@gmail.com

Research interests: AI safety & alignment, pluralistic value alignment, RLHF and post-training of LLMs, reinforcement learning, social choice & game theory, causal inference.

EXPERIENCE

Meta Superintelligence Labs
Research Engineer (contract) — Nov. 2025 – present

- Post-training of frontier LLMs: reinforcement learning, supervised fine-tuning, and coding/agentic training-data generation; build agent scaffolds, run ablation studies.

University of Oxford
Postdoctoral Researcher — Dec. 2025 – present

- Contribute the social-choice theory to *EGGROLL-IPO* (with Jakob Foerster), a decentralised post-training method that aligns an LLM to a population's diverse preferences instead of a lab-authored constitution.
- Work with Chris Summerfield on Habermas Machine v2.0 (AI-mediated deliberation) and pluralistic alignment.

Google DeepMind
Student Researcher — May – Dec. 2024

- First-authored *Jackpot! Alignment as a Maximal Lottery*: "democratic" alignment via probabilistic social choice (maximal lotteries); showed Nash Learning from Human Feedback approximates them.
- First-authored *Utility-inspired Reward Transformations*: reward transformations (from economic theory and the Inada conditions) that improve RLHF training of LLMs.
- Co-authored *Soft Condorcet Optimization for Ranking of General Agents* (Best Paper, AAMAS 2025); proved the algorithm's main properties.

Constellation Research Center
Visiting Fellow, AI Safety — Berkeley, June – Sept. 2025

- Independent research on pluralistic alignment (aligning LLMs to diverse human values).

Amazon
Applied Scientist Intern — Jun.–Sep. 2022 & Jul.–Dec. 2023

- Built multimodal CLIP-style models for text-guided image retrieval and transformer-based recommenders for Amazon Music ("Learning to Rank").

Banco de España (Bank of Spain)
Visiting Researcher, Economics — Summers 2020 & 2021

- Solved macroeconomic models with deep learning; built credit-default models and studied explainable-AI ethics.

EDUCATION

London School of Economics — PhD, Economics — 2019 – 2025
Distinction in all first-year courses.

University College London — Intercollegiate Student — 2021 – 2022
Graduate machine-learning courses at the Gatsby Computational Neuroscience Unit.

London School of Economics — MSc, Economics 2018 – 2019
Best student, EC476 (Contract Theory).

Universitat de Barcelona — BSc Mathematics & BA Business 2013 – 2020
Dual degree; top student; twelve courses graded with honors (top 5%).

University of Pennsylvania — Exchange Program 2016 – 2017
GPA 3.9/4.0; exchange awarded to two UB students on academic merit.

SKILLS

Machine learning: LLM post-training (RLHF, supervised fine-tuning, RL fine-tuning), reward modelling, model-spec / constitution evaluations, ML experiment design.
Frameworks: Python — PyTorch, TensorFlow, Transformers (Hugging Face).
Foundations: social choice & game theory, causal inference, deep learning.

PUBLICATIONS

- **Maura-Rivero, R.-R.**, Lanctot, M., Visin, F., & Larson, K. (2025). "Jackpot! Alignment as a Maximal Lottery." arXiv:2501.19266; SC4AI workshop, 2025.
- Lamerton, A., Sarkar, B., **Maura-Rivero, R.-R.**, & Foerster, J. (2026). "EGGROLL-IPO: Pluralistic Alignment via Decentralised Post-Training with Population Preferences." ICML 2026 Workshop on Pluralistic Alignment.
- **Maura-Rivero, R.-R.**, Nagpal, C., Patel, R., & Visin, F. (2025). "Utility-inspired Reward Transformations Improve Reinforcement Learning Training of Language Models." arXiv:2501.06248; ALA workshop, AAMAS 2026.
- Lanctot, M., Larson, K., *et al.* (incl. **Maura-Rivero, R.-R.**) (2025). "Soft Condorcet Optimization for Ranking of General Agents." AAMAS 2025 — **Best Paper Award**. arXiv:2411.00119.
- Estevan, A., **Maura, R.**, & Valero, Á. (2023). "Quasi-Metrics for Possibility Results: Intergenerational Preferences and Continuity." *Mathematics* **11**(2), 395.
- Kireyev, P., & **Maura-Rivero, R.-R.** (2026). "From Microeconomics to AI Research: A Guide for Economists." SSRN working paper.

HONORS & SCHOLARSHIPS

- Finalist, **Facebook Research Fellowship** in Computational Economics.
- PhD fully funded (cost + stipend) by the **Rafael del Pino**, **Bank of Spain**, and **LSE** scholarships; MSc fully funded by the **La Caixa** Fellowship.

FELLOWSHIPS, SERVICE & MISC.

- Invited expert, **European Commission** Forum on Frontier AI (2026); invited lecture, **Cambridge** (MPhil Economics & Data Science, 2026).
- **Constellation** Visiting Fellow (2025, Berkeley); **Pivotal** Fellow (2025); member, **LISA** (London Initiative for Safe AI); teach reinforcement learning at **ARENA**.
- Languages: Spanish (native), English (fluent).