

Pablo Bernabeu Pérez

✉ pbernabeuperez@gmail.com | 📧 pbernabeup | 🌐 pbernabeup | 🎓 Pablo Bernabeu-Pérez

Experience



Researcher

COEFFICIENT GIVING

Remote
Feb 2026 – Present

CG-funded project building and evaluating a benchmark of distributed attacks in multi-agent coding environments.



Research Mentor

ALGOVERSE

Remote
Jul 2025 – Present

Mentoring research groups on [Debate for AI Control](#), [Spectral Clustering for SPDs](#) and Model Organisms of Shutdown Resistance.



Research Mentor

SPAR

Remote
Sep 2025 – May 2026

Co-mentored two groups conducting research on [Multi-Signal AI Control](#) and [Coordinated Sabotage Evals](#).



Course Facilitator and Mentor

BLUE DOT IMPACT

Remote
Nov 2025 – Apr 2026

Facilitated 4 groups of the [Technical AI Safety](#) course, mentored the first cohort of the [AI Safety Project Sprint](#) and reviewed 20+ sprint projects.



External Research Collaborator

ML ALIGNMENT & THEORY SCHOLARS

Remote
Jun 2025 – Apr 2026

Co-developed C^3 , an unsupervised zero-shot black-box detection method that exploits inconsistencies in how sandbaggers respond across paraphrased prompts. Our project was awarded a £200,000 [AISI Challenge Fund Grant](#).



AI Safety Research Fellow

MARS

Remote
Jul 2025 – Oct 2025

Investigated [generalization of monitoring protocols](#) trained on synthetic model organisms to emergent misbehaviors.



AI Safety Research Fellow

LASR LABS

London, UK
Feb 2025 – May 2025

Extended the AI Control framework to [CoT monitoring](#), outperforming action-only monitors across all tested models and tasks, and attack policies designed to subvert monitoring.



Research Engineer

BARCELONA SUPERCOMPUTING CENTER

Barcelona, Spain
Nov 2023 – Jan 2025

Developed [SuSy](#) (synthetic image detector), [Aloe](#) (state-of-the-art healthcare LLM), and [CareQA](#) (multilingual QA medical benchmark) as a member of the [HPAI](#) research group.



Machine Learning Intern

SERIMAG

Barcelona, Spain
Feb 2023 – Jul 2023

Trained models for document classification, segmentation and key data extraction, including a CNN that halved the error rate of the previous document classifier.



Machine Learning Intern

FORJA IDEAS

Valencia, Spain
May 2022 – Dec 2022

Designed and developed a time series prediction model for investment funds and decision-making tool to identify and recommend the best investment opportunities.

Education



M.Sc. Artificial Intelligence

POLYTECHNIC UNIVERSITY OF CATALONIA

Average Mark: 8.84/10

Barcelona, Spain

Sep 2022 – Jun 2024



B.Sc. Computer Science

POLYTECHNIC UNIVERSITY OF VALENCIA

Average Mark: 8.78/10 (6 Subjects with Honors) · High Academic Performance Group · Co-founder of [CP-UPV](#)

Valencia, Spain

Sep 2018 – Jun 2022

Publications

Same Question, Different Lies: Cross-Context Consistency (C^3) for Black-Box Sandbagging Detection

SECOND AUTHOR

ICML - 2026

Cross-Layer Clustering for Stochastic Parameter Decomposition

SENIOR AUTHOR

LIT Workshop (ICLR) - 2026

CoT Red-Handed: Stress Testing Chain-of-Thought Monitoring

JOINT FIRST AUTHOR

NeurIPS - 2025

Present and Future Generalization of Synthetic Image Detectors

FIRST AUTHOR

ECML PKDD - 2025

Automatic Evaluation of Healthcare LLMs Beyond Question Answering

CONTRIBUTOR

NAACL - 2025

Enhancing Monitoring in AI Control through Trusted Debate

SENIOR AUTHOR

ResponsibleFM Workshop (NeurIPS) - 2025

The Aloe Family Recipe for Open and Specialized Healthcare LLMs

CONTRIBUTOR

npj Digital Medicine - 2025

Aloe: A Family of Fine-tuned Open Healthcare LLMs

CONTRIBUTOR

Preprint - 2024

Awards and Honors

Invited Visiting Researcher

CONSTELLATION

Berkeley, United States - 2025

Best Applied AI Master's Thesis Award

CATALAN ASSOCIATION OF ARTIFICIAL INTELLIGENCE

Barcelona, Spain - 2024

Best Poster Award

18TH RES USERS CONFERENCE

Málaga, Spain - 2024

1st Place

ADA BYRON PROGRAMMING CONTEST

Valencia, Spain - 2020, 2022

Honor Mention

SOUTHWESTERN EUROPE REGIONAL CONTEST

Milan, Italy - 2021

Additional Education

Mediterranean Machine Learning Summer School

FESB, UNIVERSITY OF SPLIT *Split, Croatia - Sep 2025*

Human-aligned AI Summer School

ACS RESEARCH & CTS, CHARLES UNIVERSITY *Prague, Czech Republic - Jul 2025*

AI Safety Fundamentals: Alignment

BLUEDOT IMPACT *Online - Oct 2024 – Feb 2025*

Skills

Programming	Python (PyTorch, PyTorch Lightning)
Tools	git · HuggingFace · W&B · $\text{\LaTeX}$
Languages	Spanish (Native) · English (C2) · French (A2)