Francisco Girbal Eiras

francisco.girbal@gmail.com · fgirbal.github.io Github · Google Scholar

About Me

I am currently an ML Research Scientist at Dynamo AI working on compliance and safety evaluations of generative AI systems, with a focus on large language models. Prior to this, I did my PhD on the topic of Trustworthy Machine Learning at the University of Oxford under the supervision of Philip Torr, Adel Bibi and M. Pawan Kumar (Google DeepMind).

Industry Experience

Dynamo AI, *ML Research Scientist* Nov. 2024 – Present
Working on next-generation compliance and safety evaluations of generative AI, particularly focusing on large language models.

Five, *Research Scientist Intern* Dec. 2022 – Jun. 2023
Worked on efficient methods to perform zero-shot and weakly-supervised referring image segmentation (i.e., segmenting an object in an image that is referred in a natural language sentence), achieving new state-of-the-art performance in the field.

Five, *Research Scientist Intern* Jun. 2021 – Sep. 2021
Extended the certified robustness technique of randomized smoothing from isotropic ℓ_p balls to *anisotropic* certificates through a simplified Lipschitz analysis-based framework.

Five, *Research Engineer* Sep. 2018 – Sep. 2020

- Led the development of safe and scalable optimization-based motion planning algorithms, working in a team with research scientists and software engineers.
- Published and presented research work developed at top tier conferences and journals within the robotics community, as well as to non-technical audiences.
- Wrote and reviewed research and development code, ensuring CI with other tools within the company.

Institute for Systems and Robotics, *Graduate Research Assistant* Apr. 2017 – Sep. 2017
Developed new methods to perform pose estimation using vanishing points in general (central and non-central) omnidirectional cameras leading to a CVPR 2018 paper.

Education

University of Oxford, Oxford, UK Oct. 2020 – Oct. 2024
DPhil (*PhD*), *Engineering Science*, AIMS
Supervisors: Prof. Philip H.S. Torr, Dr. Adel Bibi, Dr. M. Pawan Kumar (Google DeepMind)

University of Oxford, Oxford, UK Oct. 2017 – Sep. 2018
MSc, Computer Science
Grade: Distinction

EPFL, Lausanne, Switzerland Sep. 2016 – Feb. 2017
Student Exchange
GPA 5.75/6

Técnico Lisboa, Lisbon, Portugal Sep. 2013 – Jul. 2016
BSc, Electrical and Computer Engineering
GPA 18/20; top 2% of class

Selected Publications

F Eiras, E. Zemour, E. Lin, V. Mugunthan, *Know Thy Judge: On the Robustness Meta-Evaluation of LLM Safety Judges*, I Can't Believe it's not Better Workshop, International Conference on Learning Representations (ICLR), 2025

F Eiras, A Petrov, PHS Torr, MP Kumar, A Bibi, *Do as I do (Safely): Mitigating Task-Specific Fine-tuning Risks in Large Language Models*, International Conference on Learning Representations (ICLR), 2025

F Eiras, R Bunel, K Dvijotham, A Bibi, PHS Torr, MP Kumar, *Efficient Error Certification for Physics-Informed Neural Networks*, International Conference on Machine Learning (ICML), 2024

F Eiras, A Petrov, B Vidgen, [...], T Darrell, Y Lee, J Foerster, *Near to Mid-term Risks and Opportunities of Open-Source Generative AI*, International Conference on Machine Learning (ICML, Position Paper), 2024 [**Oral**]

F Eiras, K Oksuz, A Bibi, PHS Torr, PK Dokania, *Segment, Select, Correct: a Framework for Weakly-Supervised Referring Segmentation*, European Conference on Computer Vision (ECCV) Workshop Proceedings, 2024

F Eiras, R Bunel, K Dvijotham, A Bibi, PHS Torr, MP Kumar, *Provably Correct Physics-Informed Neural Networks*, 2nd Workshop on Formal Verification of Machine Learning, International Conference on Machine Learning (ICML), 2023 [**Outstanding Paper Award**]

A Petrov, **F Eiras**, A Sanyal, PHS Torr, A Bibi, *Certifying Ensembles: A General Certification Theory with $\mathcal{S}$-Lipschitzness*, International Conference on Machine Learning (ICML), 2023

F Eiras, M Alfarra, MP Kumar, PHS Torr, PK Dokania, B Ghanem, A Bibi, ANCER: *Anisotropic certification via sample-wise volume maximization*, Transaction of Machine Learning Research (TMLR), 2022

H Pulver, **F Eiras**, L Carozza, M Hawasly, S Albrecht, S Ramamoorthy, *PILOT: Efficient Planning by Imitation Learning and Optimisation for Safe Autonomous Driving*, International Conference on Intelligent Robots and Systems (IROS), 2021

F Eiras, M Hawasly, SV Albrecht, S Ramamoorthy, *A Two-Stage Optimization-based Motion Planner for Safe Urban Driving*, Transaction on Robotics (T-RO), 2021

SV Albrecht, C Brewitt, J Wilhelm, B Gyevnar, **F Eiras**, M Dobre, S Ramamoorthy, *Interpretable Goal-based Prediction and Planning for Autonomous Driving*, International Conference on Robotics and Automation (ICRA), 2021

Honors & Awards

Outstanding Paper Award (by European Lighthouse on Secure and Safe AI), *2nd Workshop on Formal Verification of Machine Learning @ ICML*, 2023

Honorable Mention – Using Computer Vision for Social Good, *LauzHack*, 2018

Graduate Research Fellowship, *Institute for Systems and Robotics*, 2017

Swiss-European Mobility Studentship, 2016

Undergraduate Research Fellowship, *Institute for Systems and Robotics*, 2016

Undergraduate Academic Excellency Award, *Técnico Lisboa*, 2013 – 2016

Invited Talks

On Fine-Tuning Risks in Closed Large Language Models
OxAI (Oxford AI Society) Mini-Conference, Feb. 2024

Towards Certified Machine Learning
Columbia University, New York University, University of California, Berkeley, Stanford University, Jul. 2023

Provably Correct Physics-Informed Neural Networks
UK AI Fellows Conference (Turing Institute), May 2023

Research Interests

LLM Safety · Certified Machine Learning · Optimization · Multimodal and Self-Supervised Learning · Adversarial Robustness · Formal Methods · Robotics

Programming/ Frameworks

Python · C/C++ · PyTorch · Tensorflow · WandB · transformers · git · CI/CD · AWS · Docker · React · HTML+CSS · Javascript · flask · POV-Ray