Javier Rando

AI Safety Researcher. Figuring out what can go wrong when we deploy AI in real-world applications.

Email: javier.rando@ai.ethz.ch
Homepage: https://javirando.com

Current Position

Doctorate in Computer Science | ETH Zürich | Sep. 2023 - ongoing

Advised by Prof. Florian Tramèr and Prof. Mrinmaya Sachan.

Awarded with the ETH AI Center Doctoral Fellowship 🏆.

Topics: AI Safety, Language Models, Red Teaming.

Experience

OpenAI Red Teaming Network | Independent Contractor | 2024 – ongoing

Participated in OpenAI led red teaming efforts to assess the risks and safety profile of OpenAI models and systems.

New York University | Visiting Researcher | 2022 – 2023

Research in the CILVR under the supervision of Prof. He He.

Topic: Language Models truthfulness.

Funded by a Long-Term Future Fund grant.

EXPAI | Co-Founder and CTO | 2020 – 2023

Explainable AI startup. Provides a tool to understand AI predictions to boost efficiency and trust.

In charge of corporate stragety, innovation and product development.

Telefonica | Data Science and Engineering Intern | 2020

Global department developing ML models for different use cases. Contributions to corporate Business Intelligence solutions.

Pompeu Fabra University | Research Assistant | 2019

Supervised by Prof. Carlos Castillo (Web Science and Social Computing Research Group) and Valerio Lorini (Joint Research Centre, European Comission).

Project funded with a Maria de Maetzu award for talented undergraduate students.

Dataset creation for flood detection in social media.

Study of potential bias that Wikipedia may exhibit when reporting natural disasters. Paper entitled Uneven Coverage of Natural Disasters in Wikipedia: The Case of Floods published at ISCRAM 2020.

Education

MSc in Computer Science | ETH Zürich | 2021 – 2023

Major in Machine Intelligence.

Research projects: Language Models for password modeling (Prof. Fernando Perez-Cruz) and poisoning RLHF (Prof. Florian Tramèr).

Funded by an Open Philanthropy grant for individuals working to improve long-term future

BSc in Mathematical Engineering in Data Science | Pompeu Fabra University | 2017 – 2021

First in class. Best Academic Record (9.1/10) 🏆.

4-year degree in English.

Year abroad in the **Technical University of Munich (TUM)**.

Publications

Pre-prints

2023. Nitish Joshi, Javier Rando, Abulhair Saparov, Najoung Kim, He He. *Personas as a way to Model Truthfulness in Language Models*.

2023. Rusheb Shah, Quentin Feuillade–Montixi, Soroush Pour, Arush Tagade, Javier Rando. *Jailbreaking Language Models at Scale via Persona Modulation*.

Proceedings

2024. Javier Rando, Florian Tramèr. *Universal Jailbreak Backdoors from Poisoned Human Feedback*. ICLR. 🏆 **2nd Prize at the AI Safety Prize Competition.** 🏆.

2023. Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, et al. *Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback*. Transactions on Machine Learning Research (TMLR).

2023. Javier Rando, Fernando Perez-Cruz, Briland Hitaj. *PassGPT: Password Modeling and (Guided) Generation with Large Language Models*. ESORICS.

2022. Edoardo Mosca, Shreyash Agarwal, Javier Rando, Georg Groh. *"That Is a Suspicious Reaction!": Interpreting Logits Variation to Detect NLP Adversarial Attacks*. ACL.

2020. Valerio Lorini, Javier Rando, Diego Saez-Trumper, Carlos Castillo. *Uneven Coverage of Natural Disasters in Wikipedia: The Case of Floods*. ISCRAM 2020 Conference Proceedings.

Workshops

2022. Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, Florian Tramèr. *Red-teaming the Stable Diffusion Safety Filter*. NeurIPS ML Safety Workshop. 🏆 **Best Paper Award** 🏆.

2022. Javier Rando, Alexander Theus, Rita Sevastjanova, Menna El-Assady. *How is Real-World Gender Bias Reflected in Language Models?* IEEE VIS VISxAI Workshop.

2022. Javier Rando, Thomas Baumann, Nasib Naimi, Max Mathys (2022). *Exploring Adversarial Attacks and Defenses in Vision Transformers trained with DINO*. ICML Adversarial ML Frontiers Workshop.

Teaching

2023 **ETH Zurich**. Teaching assistant for Information Security Lab. Module on LLM safety.

2023 **NYU AI School**. Teaching assistant for introductory lab courses to AI.

2021 **Pompeu Fabra University**. Guest lectures (4h). Visual Analytics undergraduate course.

2021 **Pompeu Fabra University**. Guest lecture (2h). Entrepreneurship undergraduate course.

Service

Organizing comittee

2023 **Competition organizer at SaTML 2023**. *Large Language Models CapturetheFlag* and *Find the Trojan: Universal Backdoor Detection in Aligned Large Language Models*

Peer reviewer

2024 **ICML**. Main conference track.

2023 **Socially Responsible Language Modelling Research (SoLaR) Workshop at NeurIPS 2023**.

Awards and Grants

2023 **ETH AI Center Doctoral Fellowship**.

2022 **Best Paper Award** at *NeurIPS ML Safety Workshop*.

2022 **Runner up** at *Junction 2022 Hackathon Helsinki. Challenge: Crypto Trading Helper*.

2022 **Junior Researcher**. *Future of Life Institute*.

2022 **Research grant** ($30K) to join NYU as Visiting Researcher. *Long-Term Future Fund*.

2021 **Early career scholarship** for MSc funding ($30K). *Open Philanthropy Project*.

2021 **Best Academic Record Award**. BSc Mathematical Engineering in Data Science (9.1/10). *Pompeu Fabra University*.

2019 **Collaboration scholarship at units of excellence "María De Maeztu"**. Research assistant funding at Pompeu Fabra University. *Spanish Science Ministry*.

2020 **TALENTUM internship Telefonica**. Internship to join an applied research team. *Telefonica*.

2017 **Best Academic Record** (10/10). *High School*.