

Alejandro Acelas

[linkedin.com/in/alejandro-acelas](https://www.linkedin.com/in/alejandro-acelas) • github.com/AlejoAcelas • alejoacelas.github.io
alejoacelas@gmail.com • Bogotá, Colombia. • (+57) 3166803746

EDUCATION

B.S. Economics

Universidad de Los Andes. GPA: 3.77/4.00 (Cum Laude)

Aug 2020 – Nov 2023

Bogotá, Colombia

- Relevant Coursework: Topics in Machine Learning, Data Analysis in R, Statistics (Honors).

EXPERIENCE

Machine Learning Engineer

Independent Contractor

Aug 2023 – Oct 2023

Remote

- Reverse-engineered the algorithms learned through gradient descent by transformer models trained on simple algorithmic tasks, for a project commissioned by Marius Hobbhahn.

Research Fellow

Swiss Existential Risk Initiative

Jun 2023 – Aug 2023

Geneva, Switzerland

- Designed a set of interpretability challenges on Transformer models. The challenges included repairing an ablated circuit, analyzing out-of-context behavior and detecting backdoor inputs
- Replicated key findings from Anthropic's 'Toy Models of Superposition' paper and explored using Gaussian perturbations to identify the features most relevant to a model's loss

Research Assistant

Under Tomás Rodríguez Barraquer at Universidad de Los Andes

Sep 2021 – Jun 2022

Bogotá, Colombia

- Analyzed a corpus of millions of tweets to investigate the polarization of discourse on social media during social protests
- Developed SIS contagion models in Python, incorporating evolutionary game theory to study how social norms affect personal protective behaviors and health outcomes in pandemics

HONORS

3 rd Place National Academic Economics Contest	2023
Honorable Mention International Mathematical Olympiad	2017
5 th Place Colombian Mathematical Olympiad (2017)	2017
1 st Place Mathematical Olympiad at Universidad Industrial de Santander	2013, 2015, 2017

MISCELLANEOUS

Alignment Research Engineering Accelerator

Participant

Apr 2023 – Jun 2023

Implemented algorithms and applications covering the foundations of ML for AI Safety, with especial emphasis on mechanistic interpretability, reinforcement learning, and training at scale.

Effective Altruism Uniandes

Leader and cofounder

Jun 2021 – Feb 2023

Organized events, mentored students and facilitated in-person programs on topics related to Effective Altruism and high-impact careers.

SKILLS

Programming Languages: Python, R, Stata, SQL, Java

Tools: Pytorch, Numpy, Pandas, Matplotlib, PySpark, AWS Services (Glue, SageMaker), ggplot2, dplyr

Languages: English (TOEFL 112/120), Spanish