

Ayrton San Joaquin

Every colored text is a link.

🔬 **SCIENTIST & SOFTWARE ENGINEER, TRUSTWORTHY AI** | ✍️ **WRITER**

✉️ ayrton@aya.yale.edu | 📍 Singapore | 🌐 [ajsanjoaquin](https://github.com/ajsanjoaquin) | 🌐 [ajsanjoaquin](https://www.linkedin.com/in/ajsanjoaquin/) | 📌 [Values](#)

Education

Yale-NUS College

BSC. (HONORS) IN DATA SCIENCE, MINOR IN PHILOSOPHY. (**SCHOLAR, WITH HIGH DISTINCTION**)

Semester Abroad at the University of Copenhagen, Denmark

Singapore

August 2018 - May 2023

Experience

French National Centre for Scientific Research (CNRS)@CREATE

Singapore

AI SCIENTIST, **DESCARTES PROGRAM** ([HTTPS://DESCARTES.CNRSATCREATE.CNRS.FR/](https://DESCARTES.CNRSATCREATE.CNRS.FR/))

September 2023 - Present

- Leading a study on efficient training of Large Language Models for critical decision-making.
- Collaborating on a psychology study on human perceptions of intelligent systems in morally-relevant urban situations.
- Co-authoring an ethnographic study of AI researchers under the program to understand their assumptions in developing technologies for the smart city.

Machine Learning Safety Scholars Program, Center for AI Safety

Palo Alto, United States

SCHOLAR

June 2022 - August 2022

- Studied model failures (CV and NLP), and led research on analyzing Large Language Models (LLMs) using few-shot learning.
- Implemented various strategies in **robustness** (PGD, adversarial training), **anomaly detection** (AUROC, ViM), **calibration** (RSME, Brier scores), and **trojan attacks** (data poisoning).

Data Privacy and Trustworthy Machine Learning Lab, NUS

Singapore

UNDERGRADUATE RESEARCHER

May 2021 - March 2022

- Collaborated with Google DeepMind on privacy and adversarial machine learning research for my bachelor's thesis in a team across 4 time zones. **Published in a top security conference (ACM CCS) as the youngest and only undergraduate co-author.**

Arterys (Freelance)

San Francisco, United States

DEEP LEARNING ENGINEER

March 2020 - June 2020

- Created a COVID-19 Pneumonia classifier **4 days after pandemic declaration in collaboration with A.I. Singapore.**
- Collaborated with Arterys to **deploy the model in their platform** for use by American hospitals and researchers. Model engineer in a team of 4 across 3 time zones.

Open-Source Projects & Public Service

The AI Summit Singapore 2024

Public Communication

- Invited as a featured speaker to discuss about data privacy for Generative AI at Asia Tech x Singapore, co-hosted by the Singapore Government.

Public AI Advocacy

Field-Building

- Co-authored the foundational document of a global network of AI researchers, policymakers, and practitioners to center AI research and development in the public interest and as public infrastructure.

Project Aria Timeline Builder Workshop 2023

Contextual AI, AR

- Invited by Meta Reality Labs to design a use-case for **Project Aria**. Participated as the only bachelor's graduate among 20 graduate and postdoctoral researchers. [Video demo.](#)

Explaining Neural Networks with Meaningful Perturbations

Explainable AI, CV

- For explaining an image classifier's prediction, I implemented the algorithm described in *Explanations of Black Boxes by Meaningful Perturbation* (Fong, et. al., 2018).

Open-Source AI

DevOps

- Added new features for major machine learning projects including Pytorch, HuggingFace Transformers, and YOLOv4 (object detection model).

Publications

November 2022 Tramer, F., ..., **San Joaquin, A.**, et.al. , Truth Serum: Poisoning Machine Learning Models to Reveal Their Secrets

[ACM CCS 2022 link](#)

March 2020 **San Joaquin, A.**, *Using Deep Learning to Detect Pneumonia caused by COVID-19*

[Towards Data Science \(Editor's Choice\) link](#)

Skills

Machine Learning: Pytorch, Sickit-Learn, JAX, Keras, Tensorflow, Transformers, Langchain, NLTK, Spacy

Data: Pandas, Querying (SQL, MySQL, PySpark, MongoDB), Vector Database (Qdrant, Pinecone)

MLOps: AWS, AzureML, Snowflake, Docker, Flask, Continuous Integration, Kubernetes, MLFlow