Every colored text is a link.

Ayrton San Joaquin

Scientist & Software Engineer, Trustworthy AI | Writer

ayrton@aya.yale.edu | Singapore | ajsanjoaquin | ajsanjoaquin | Values

Education

Yale-NUS College — *Singapore*

BSc. (Honors) in Data Science, Minor in Philosophy. **(Scholar, With High Distinction)** — *August 2018 - May 2023*

Semester Abroad at the University of Copenhagen, Denmark

Experience

French National Centre for Scientific Research (CNRS)@CREATE — *Singapore*

AI Scientist, DesCartes program (https://descartes.cnrsatcreate.cnrs.fr/) — *September 2023 - Present*

- Leading a study on efficient training of Large Language Models (LLMs) 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 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 — *Information Security*

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

Public AI Advocacy — *Trustworthy AI*

• 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 — *LLMs, Contextual AI*

• 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.

Eplaining 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, HuggingFace, Langchain, NLTK, Spacy

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

MLOps: Linux, Git, GCP, AzureML, Snowflake, Docker, Flask, Continuous Integration, Kubernetes