Every colored text is a link.

Ayrton San Joaquin

Scientist, Trustworthy AI | Writer

ayrton@aya.yale.edu | Barcelona, Spain | 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

AI Standards Lab *Singapore & Spain*

Research Analyst *August 2024 - Present*

- Contributing to the Codes of Practice of the EU AI Act.

French National Centre for Scientific Research (CNRS)@CREATE
International Research Laboratory on Artificial Intelligence (IPAL- by appointment) *Singapore*

AI Scientist, DesCartes program *September 2023 - August 2024*

- Led a Franco-Singaporean team studying efficient fine-tuning of Large Language Models (LLMs). Published at EMNLP Findings 2024.
- Created and curated the Tagalog dataset of https://seaeval.github.io/, a multilingual benchmark for Southeast Asian Languages.

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.

Honors & Public Service

AI Safety Connect 2025 *Paris, France*

• Invited to present about risk management, an official side-event of the Paris AI Action Summit.

International Association for Safe and Ethical AI Conference 2025 *Paris, France*

• Invited to present about risk management at IASEAI Conference, an official side-event of the Paris AI Action Summit

The AI Summit Singapore 2024 *Singapore*

• 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 *Global*

• Co-authored a 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 *Menlo Park, United States*

• 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.

Publications

January 2025	Gipiškis, R. and **San Joaquin, A.,** et. al. , Risk Sources and Risk Management Measures in Support of Standards for General-Purpose AI Systems. *Under Review*	*link*
November 2024	**San Joaquin, A.,** et. al. , In2Core: Leveraging Influence Functions for Coreset Selection in Instruction Finetuning of Large Language Models. *EMNLP Findings 2024*	*link*
November 2022	Tramer, F., ..., **San Joaquin, A.,** et.al. , Truth Serum: Poisoning Machine Learning Models to Reveal Their Secrets. *ACM Conference on Computer and Communications Security (CCS) 2022*	*link*