Cosmo Nazzareno Santoni

Email: santonicosmo@gmail.com | cosmo.santoni@imperial.ac.uk | Tel: +44 7444 403542
Links: GitHub | Google Scholar | Personal Website | LinkedIn

Machine learning researcher specialising in sequence models, state-space architectures, and simulation-based inference for decision-making and LLM robustness and safety, with 125,000× speed-ups over traditional simulators while maintaining calibrated uncertainty estimates. Deployed in production for WHO malaria programme planning and UK government crisis response; publications include Nature and The Lancet, with work presented at NeurIPS and under work at ICML and KDD ADS.

SELECTED IMPACT

- Developed **state-space and RNN neural surrogate architectures** for large-scale agent-based models, achieving **125,000× speedups with 99.8% fidelity**, deployed **in production for WHO** malaria intervention planning and international government decision support.
- Co-authored high-impact modelling work in **Nature** and **The Lancet**, and work under submission at **ICML**
- Received **SAGE Award** from **UK Government Chief Scientific & Medical Officers** for modelling and data support enabling evidence-based COVID-19 policy.
- Released **LLM robustness and safety tooling**, including a **metacognitive study of Claude reviewed by Anthropic Trust & Safety**.
- Built the differentiable **JAX-based** 1D arterial hemodynamics **simulator and hierarchical SBI dataset generator** underpinning **our NeurIPS 2025 workshop** and **ICML 2026 manuscript** tokenised flow matching work on hierarchical inference.

PROFESSIONAL EXPERIENCE

Machine Learning Researcher / Technical Analyst – *Imperial College London, U.K., 2023 – Present*

- Led end-to-end design and research of state-space neural surrogate architectures for large-scale agent-based simulators, replacing 24-minute HPC computations with sub-10.5ms inference (>125,000× speedup, 99.8% fidelity) and cut compute from ~22–43 MWh on consumer/HPC CPUs (≈£5.9k–£11.3k) to ~0.13 kWh for ~1M scenarios.
- Architected the MINTverse production ML platform processing 100+ scenarios per second on a consumer GPU. Designed a modular neural multi-task emulation architecture with hot-swappable RNN / Mamba-2 sequence-model heads (PyTorch + CUDA + AMP), built data infrastructure and analytics over 6.89B+ data points (DuckDB, sub-second queries), implemented model serving APIs, and deployed with deterministic versioning—enabling millisecond-latency policy queries where hours were previously required.
- Built open-source modular ML infrastructure (MINTverse: segMINT, estiMINT, MINTe, MINTer, etc.) that reduced internal teams' model development and experimentation cycles from months to days.
- Partnered with WHO to translate neural emulator outputs into deployment guidance, bridging ML model performance, uncertainty calibration, and policy requirements for international malaria intervention programmes.

Machine Learning Researcher – *University of Cambridge, Department of Computer Science, UK., & German Centre for Artificial Intelligence (DFKI), DE., 2022 – 2023*

- Engineered neural ODE architectures with hard physical constraints for complex dynamical systems. Implemented physics-informed loss functions and automatic differentiation in PyTorch / Julia, achieving strong performance on irregular sparse time-series while maintaining clinical interpretability.
- Designed training and deployment infrastructure for multi-scale dynamical systems, building pipelines supporting multiple optimisation algorithms (Adam, Levenberg-Marquardt), Bayesian uncertainty quantification, and sensitivity analysis suitable for safety-critical settings.
- Drove architecture decisions for Neural Universal Differential Equations, balancing expressiveness vs interpretability. Developed 1D/2D neural ODE architectures that enabled previously intractable dynamical systems where unconstrained data-driven baselines failed physical consistency checks.

Research Assistant – *COVID-19 Real-time Modelling, Imperial College London, U.K., 2021 – 2023*

- Owned real-time Bayesian inference pipeline (particle MCMC in C++ / R) for national-scale epidemic time-series, delivering <24hr from raw data to policy recommendations with sub-hour end-to-end runtimes. Outputs directly informed UK lockdown policy during the national emergency.
- Maintained mission-critical open-source Bayesian inference libraries (sircovid, spimalot, MCState) under active use by UK SAGE during the pandemic. Managed breaking changes, backward compatibility, and emergency bug fixes while ensuring reproducibility across distributed teams.
- Engineered automated reporting infrastructure generating statistical analyses, visualisations, and forecasts for systems running continuously for 18+ months, supplying weekly briefings to UK Chief Scientific Advisors.
- Received SAGE Award for Modelling and Data Support from UK Government Chief Scientific and Medical Officers, recognising technical contributions to large-scale modelling and data pipelines for national-scale decision-making.

TECHNICAL SKILLS & LANGUAGES

- ***Languages***: *Python, R, Julia, C++ (familiar), C#, SQL, Bash (Linux)*
- ***ML / DL***: *PyTorch, JAX, JAX-CFD, MLX, Mamba2, XGBoost, Optuna, NumPy, Pandas*
- ***Data & Storage:*** *DuckDB, HDF5*
- ***HPC & Acceleration***: *CUDA, MPS, Automatic Mixed Precision (AMP); CPU / GPU clusters*
- ***Tooling:*** *Git / GitHub, CI/CD (GitHub Actions), Linux*
- ***ML focus:*** *sequence models & state-space models (Mamba-2), neural ODEs, simulation-based inference, Bayesian time-series modelling, calibration and uncertainty quantification.*

EDUCATION

- **Full-time PhD in Applied Mathematics,** *Imperial College London, U.K., January 2025 – December 2027*
  Thesis: "Accelerating Real-Time Agent-Based Simulations with State-Space Neural Surrogates and Graph Neural Networks"
- **MSc. Epidemiology (Merit),** *Imperial College London, U.K., 2020 – 2021*
- **BSc. (Hons.) Mathematics with Economics (2:1),** *Aston University, U.K., 2015 – 2019*

PUBLICATIONS & RESEARCH OUTPUT

Selected Publications:

- *Perez-Guzman, P. N., Knock, E., et al. "Epidemiological drivers of transmissibility and severity of SARS-CoV-2 in England." **Nature Communications (co-author)***
- *Imai, N., Rawson, T., et al. "Quantifying the impact of delaying the second COVID-19 vaccine dose in England: a mathematical modelling study" **The Lancet Public Health (co-author)***
- *Further Publications Under PERG including **Lancet Infectious Disease, Lancet Global Health & Lancet Microbe***

Manuscripts under Preparation/Review:

- *Charles, G, et al. "Tokenised Flow Matching for Hierarchical Simulation Based Inference"* **International Conference on Machine Learning, ICML,** *in preparation, 2026* ***(co-author)***
- *Santoni, N. C., et al. "State-Space Neural Emulators for Real-Time Policy Decision Support"* ***ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD****, in preparation, 2026.* ***(first-author)***

Production ML Systems (World Health Organisation)

- *MINTverse: production orchestration (R/Python) delivering neural emulation for WHO operational decision-support.*
- *RADAR (in-dev): state-space error calibration for time-series emulators (private; results on request)*

LLMs, alignment & safety

- *mamba2-jax: A pure JAX/Flax implementation of Mamba-2 for language modeling and time series forecasting.*
- *mamba2-triton-guard: Lightweight patcher that stubs Triton and guards version checks so mamba2_torch imports cleanly on macOS M1–M5 (CPU / MPS), enabling SSM experimentation without GPU-only constraints.*
- *mamba2_torch contribution (PR under review): Added import guard for MPS, gated fast-path, and fixed a padding edge case to improve stability on Apple silicon and avoid silent shape errors.*
- *Metacognitive / Continuity Prompting of Claude - independent technical report reviewed by Anthropic Trust & Safety, providing a systematic behavioural analysis of non-uniform helpfulness and a proposed alignment evaluation framework.*
- *ALETHIA + KAIROS: metacognitive control and selective acceptance framework for LMs, with full audit artefacts and open-source implementation.*

Additional Open-source libraries

- *Epireview: production data extraction and automated figures / tables for WHO PERG; used by international teams*
- *Sircovid, Spimalot, MCState: UK COVID-19 Bayesian toolkit (adaptive PMCMC, inference, forecasting).*

CONFERENCES & PRESENTATIONS

- **Tokenised Flow Matching for Hierarchical Simulation Based Inference**, *NeurIPS Workshop on Simulation-Based Inference, San Diego, USA, December 2025 (co-author; presented by Charles, G.)*
- **Epireview: Hands-on Workshop for Public Health & Epidemiology Researchers,** *Infectious Disease Modelling Conference, Bangkok, Thailand, November 2024*
- **Applying Neural Network Emulation to Assess the Impact of Pyrethroid-Pyrrole Bed Nets on Malaria in Africa,** *9th International Conference on Infectious Disease Dynamics, Bologna, Italy, November 2023*

ACADEMIC SERVICE & VOLUNTARY WORK

- **Curator, Amphibian & Malaria Collections**, *Museum of Life Sciences, King's College London*, U.K., 2025 –
- **Departmental MRC GIDA Seminar Series Co-Organiser**, *Imperial College London, U.K.*, 2023 –