Cosmo Nazzareno Santoni
Email: santonicosmo@gmail.com | cosmo.santoni@imperial.ac.uk | Tel: +44 7444 403542
Links: GitHub | Google Scholar | Personal Websites | LinkedIn
Location: London, UK (Available to relocate)

Machine learning researcher, engineer and PhD candidate in Applied Mathematics at Imperial College London specialising in sequence models, state-space architectures, and simulation-based inference for decision-making and large language models (LLMs), LLM robustness and safety, with 125,000× speed-ups over traditional simulators while maintaining calibrated uncertainty estimates. Deployed in production for World Health Organisation (WHO) malaria programme planning and UK government crisis response; publications include Nature and The Lancet with work presented at a NeurIPS 2025 workshop and manuscripts under review at top-tier ML conferences 2026 and TMLR. Creator of mamba2-jax, a JAX/Flax Mamba-2 implementation merged by the Bonsai JAX team into Google's official jax-ml/bonsai models repository.

SELECTED IMPACT

- Contributed the **Mamba-2** (JAX/NNX) implementation to **Google's jax-ml/bonsai model zoo**, including pretrained-weight loading and state-space caching (up to **28×** inference speedup). Ongoing contributor; co-authoring a Google Open Source Blog post with the JAX Bonsai team.
- Developed **state-space and RNN neural surrogate architectures** for large-scale agent-based models, achieving **~125,000× speedups** against reference agent-based simulator, deployed **in production for WHO** malaria intervention planning and international government decision support.
- Received **SAGE Award** from **Sir Patrick Vallance & Professor Chris Whitty, UK Government Chief Scientific & Medical Officers** for modelling and data support enabling evidence-based COVID-19 policy.
- Released **LLM safety tooling prototypes**, including a metacognitive study of Claude; transcripts shared with **Anthropic Trust & Safety**.

PROFESSIONAL EXPERIENCE

Machine Learning Research Software Engineer – *Imperial College London, U.K., 2023 – Present*

- Led end-to-end design and research of state-space neural surrogate architectures for large-scale contextual agent-based simulators, replacing 24-minute HPC computations with sub-10.5ms inference (>125,000× speedup; calibrated uncertainty preserved; near-parity on held-out scenarios) and cut compute from ~22–43 MWh on consumer HPCs (≈£5.9k–£11.3k) to ~0.13 kWh for ~1M scenarios.
- Lead MINTverse production; MLOps / 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 synthetic data generation and analytics infrastructure over 6.89B+ data points (DuckDB, sub-second queries), and deployed model serving APIs with deterministic versioning—enabling millisecond-latency policy queries where hours would have previously been required.
- Built open-source modular ML infrastructure that reduced teams' model development cycles from months to days.
- Partnered with WHO to translate neural emulator outputs into deployment, bridging ML model performance 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 inference pipeline for national-scale epidemic time-series, delivering <24hr from raw data to policy recommendations. Outputs directly informed UK lockdown policy during the national emergency.
- Maintained open-source inference libraries 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.

TECHNICAL SKILLS & LANGUAGES

- ***Languages***: *Python, R, Julia, C++ (familiar), SQL, Bash (Linux)*
- ***ML / DL:*** *JAX, PyTorch, JAX-CFD,* ***Transformers****, Mamba2, RNNs/LSTMs, XGBoost, Optuna, NumPy, Pandas*
- ***Data & Storage:*** *DuckDB, HDF5*
- ***HPC & Acceleration***: *CUDA, MPS, Automatic Mixed Precision (AMP); CPU/GPU clusters; Google Cloud TPUs / GCP (familiar); Triton kernels (familiar), distributed training (DeepSpeed, ZeRO, TP/PP/DDP – familiar)*
- ***Tooling:*** *Git / GitHub, CI/CD (GitHub Actions), Linux, Docker (deployment exposure)*
- ***Research focus:*** *RNNs, state-space models (Mamba-2), Transformers and large language models (LLMs), generative modelling (flow matching), neural ODEs, simulation-based inference (SBI), Bayesian optimisation and modelling, continual learning, conformal prediction, calibration and uncertainty quantification (UQ).*

EDUCATION

- **PhD in Applied Mathematics,** *Imperial College London, U.K., January 2025 – December 2027*
Thesis: "Towards State Space and Continual Learning Models for Large-Scale Agent-Based Simulators"
- **MSc. Epidemiology (Merit),** *Imperial College London, U.K., 2020 – 2021*
- **BSc. (Hons.) Mathematics with Economics,** *Aston University, U.K., 2015 – 2019*

SYSTEMS, OPEN SOURCE & RESEARCH SOFTWARE

Production ML Systems & Flagship Libraries

- *mamba2-jax: JAX/Flax Mamba-2 implementation. Core architecture (PR #103) with state-space caching (PR #131), merged into Google's jax-ml/bonsai; ongoing collaboration with the Bonsai team.*
- *MINTverse: production orchestration (R/Python) delivering neural emulation for WHO operational decision-support.*

Robustness, Uncertainty Quantification, LLMs

- *Metacognitive Prompting of Claude - systematic behavioural analysis of non-uniform helpfulness; shared transcripts with Anthropic Trust & Safety.*
- *RADAR: risk-aware UQ and calibration for SSMs, learning residual-based risk scores and adaptive conformal prediction under distribution shift, designed for time-series, LLM token, sequence, and reasoning-head outputs.*

Open-source libraries & tooling

- *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 triton-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.*
- *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).*

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)***
- *Multiple publications under PERG including* ***Lancet Infectious Disease, Lancet Global Health & Lancet Microbe***

Manuscripts under Preparation/Review:

- *Charles, G., Santoni, N. C., et al. "Tokenised Flow Matching for Hierarchical Simulation Based Inference"*, **Manuscript under review (venue withheld due to double-blind policy; co-author)**
- *Santoni, N. C., et al. "State-Space Neural Emulators for Real-Time Policy Decision Support"* ***TMLR, 2026****.* ***(target venue; first-author)***
- *Santoni, N. C., Schwarz, J. R., et al. "Freeze-Thaw Bayesian Optimisation for Accelerated Data Mixture Learning", (****In preparation; first-author)***

Conferences, Workshops & Presentations:

- **Tokenised Flow Matching for Hierarchical Simulation-Based Inference**, *NeurIPS 2025 Workshop on Frontiers in Probabilistic Inference: Sampling Meets Learning*, *San Diego, USA*

ACADEMIC SERVICE & VOLUNTARY WORK

- **Foundational Machine Learning Research Working Group Co-chair**, Imperial College London, U.K., 2026 –
- **Curator, Amphibian & Malaria Collections**, *Museum of Life Sciences, King's College London*, U.K., 2025 –
- **Departmental Seminar Series Co-Organiser**, *Imperial College London, U.K.*, 2023 – 2025