Plan Overview

A Data Management Plan created using DMPonline

Title: FLF Data Management

Creator: Pietro Ferraro

Principal Investigator: Pietro Ferraro

Affiliation: Imperial College London

Funder: UKRI Future Leaders Fellowships

Template: UKRI Future Leaders Fellowships

Project abstract:

A foundational mathematical and computational fellowship of decentralised agentic AI, combining theoretical analysis with large-scale agent-based simulations and hardware validation.

ID: 206129

Start date: 01-09-2027

End date: 31-08-2034

Last modified: 09-06-2026

Copyright information:

The above plan creator(s) have agreed that others may use as much of the text of this plan as they would like in their own plans, and customise it as necessary. You do not need to credit the creator(s) as the source of the language used, but using any of the plan's text does not imply that the creator(s) endorse, or have any relationship to, your project or proposal

FLF Data Management - Initial DMP

Management and sharing data

How will you manage and share data collected or acquired through the proposed research?

Question not answered.

FLF Data Management - Detailed DMP

0. Proposal name

0. Enter the proposal name

Foundations of Decentralised Artificial Intelligence

1. Description of the data

1.1 Type of study

A foundational mathematical and computational fellowship of decentralised agentic AI, combining theoretical analysis with large-scale agent-based simulations and hardware validation. The study does not involve human subjects, identifiable personal data, animals or biological samples.

1.2 Types of data

Quantitative only, generated from: (i) computational simulations (agent state trajectories, control-performance metrics, safety-certificate violation rates, population-level stability metrics); (ii) controlled hardware experiments on an inverted-pendulum testbed (sensor readings and control outputs); (iii) benchmark outputs from safe-AI gym environments, SUMO smart-city scenarios, and the Responsible AI Institute agent-interaction sandbox; (iv) call/response logs from frontier LLM APIs (OpenAI, Anthropic, etc.) used for benchmarking agent populations. No qualitative, clinical, medical, genotypic, tissue or administrative records are involved.

1.3 Format and scale of the data

File formats: HDF5 for high-dimensional simulation trajectories; Parquet and CSV for tabular benchmark outputs; JSON and YAML for experiment configurations and metadata; PyTorch / JAX checkpoint files for learned model weights; ONNX for cross-framework model interchange where applicable; LaTeX and PDF for written outputs. Software: open-source Python stack (NumPy, SciPy, PyTorch, JAX, Stable-Baselines3, SUMO), formal-methods toolchains (CasADi, dReal), with all dependencies pinned via pyproject.toml or requirements.txt. All chosen formats and tools are openly readable, widely adopted in the community, and support long-term validity.

2. Data collection / generation

2.1 Methodologies for data collection / generation

Data are generated programmatically through reproducible simulation pipelines and through controlled experiments. Code is version-controlled in Git (private GitHub repositories during development; public on paper acceptance). The project follows the FAIR data principles (go-fair.org/fair-principles). Where domain standards exist they will be used: SUMO XML standards for smart-city scenarios, standard safe-AI benchmark interfaces, and ONNX where cross-framework model interoperability is needed.

2.2 Data quality and standards

Reproducibility is enforced through: (i) seed-controlled stochastic simulations with seeds and configurations stored alongside outputs; (ii) version-pinned dependency environments via Docker / Singularity containers for major experiments; (iii) automated continuous-integration test suites for code; (iv) repeat runs and ablation studies for any quantitative claim that goes into a paper; (v) a standardised JSON schema describing each experiment, co-located with every output; and (vi) external peer review through publication at top-tier venues (IEEE TAC, NeurIPS, ICML, AAAI, IEEE Internet of Things Journal).

3. Data management, documentation and curation

3.1 Managing, storing and curating data

Active data are held on Imperial's Research Computing Service (RCS) and on the Imperial Research Data Facility (RDF) Active, both of which provide automated backup, snapshotting and access control. Code is stored in version-controlled Git repositories. Project documentation and metadata are held within Imperial's OneDrive / SharePoint infrastructure. Curation follows the FAIR data principles (go-fair.org/fair-principles) and Imperial's Research Data Management guidance (imperial.ac.uk/research-and-innovation/.../research-data-management).

3.2 Metadata standards and data documentation

Each released dataset is accompanied by: (i) a README describing purpose, generation method, fields and units; (ii) a JSON schema describing variables, types and ranges; (iii) the seeds, hyperparameters and software/hardware environments used; (iv) provenance information linking each output to the Git commit hash of the code that produced it; (v) where applicable, model cards and dataset cards following established machine-learning community conventions. Metadata is human-readable and machine-readable.

3.3 Data preservation strategy and standards

Data underpinning published findings will be deposited in Imperial's institutional repository (Spiral, spiral.imperial.ac.uk) and/or Zenodo (zenodo.org), each of which assigns a persistent DOI and provides long-term curation. Retention: at least 10 years from publication, in line with UKRI's research-data policy (UKRI Research Data Policy). Intermediate exploratory simulation outputs not used in publications may not be retained beyond the active phase of the project, on cost-effectiveness grounds.

4. Data security and confidentiality of potentially disclosive information

4.1 Formal information/data security standards

Imperial College London operates within an information-security framework aligned with ISO/IEC 27001 principles.

Imperial ICT Services ISO 27001 certificate number: 210676.

OneDrive for Business is authenticated to meet the College's data security and resilience standards and to comply with ISO27001, HIPAA, FISMA, the US-EU Safe Harbor framework, and the EU Data Protection Directive model clauses.

4.2 Main risks to data security

The fellowship does not collect or process identifiable personal data, and the principal scientific activity involves no human participants. Risks are therefore low. Where partner-provided datasets are used (e.g. from Verismart deployments and the Responsible AI Institute agent-interaction sandbox), any such data are pseudonymised or aggregated by the providing partner before they reach the fellowship team, and are governed by Data Sharing Agreements. Access to those data is restricted to named team members under those agreements. Imperial Research Data Facility (RDF) Active and Research Computing Service provide encryption at rest and in transit, role-based access control, access logging, and audited user permissions.

5. Data sharing and access

5.1 Suitability for sharing

Yes. The data are predominantly simulation-generated and methodological in nature, and openly sharing them accelerates community uptake, reproducibility and downstream research. Open-source release of code, datasets and reference implementations is a core deliverable of the fellowship.

5.2 Discovery by potential users of the research/innovation data

Released datasets and code will carry persistent DOIs from Imperial Spiral and/or Zenodo and will be discoverable through: (i) the relevant peer-reviewed publication; (ii) the DecAIL website; (iii) Pietro Ferraro's GitHub; and (iv) standard ML community discoverability surfaces (Papers with Code; HuggingFace where applicable). The publishing repositories (Spiral and Zenodo) are openly accessible to anyone with internet access. The fellowship's data-sharing policy will be published on the DecAIL website.

5.3 Governance of access

For openly published datasets and code (the great majority of fellowship outputs), no per-request decision is required: access is open. For any partner-restricted data, Dr Pietro Ferraro (PI) is the decision-maker, in consultation with the data-providing partner under the governing Data Sharing Agreement. Long-term archived data in Imperial Spiral is subject to Imperial's standard access policies. Datasets will be deposited in Imperial Spiral and/or Zenodo as appropriate; community-specific repositories (e.g. HuggingFace) will additionally be used where they aid discoverability.

5.4 The study team's exclusive use of the data

Code and reproducibility artefacts are released openly at the time of publication of the corresponding peer-reviewed paper. Open datasets associated with publications are released at the same time, or sooner where this is in the project's interest. The exclusive-use period is limited to the period from data generation to first peer-reviewed publication (typically six to twelve months), after which the data are openly available. For partner-restricted data, sharing timescales follow the governing Data Sharing Agreement.

5.5 Restrictions or delays to sharing, with planned actions to limit such restrictions

The fellowship does not involve human participants, so consent-based restrictions do not arise. Principal restrictions arise from partner-shared data (Verismart, the Responsible AI Institute sandbox), where pseudonymisation, aggregation and Data Sharing Agreements limit access.

5.6 Regulation of responsibilities of users

For openly published data and code, external users are bound by the relevant open-source licence (MIT or Apache 2.0 for code; CC BY 4.0 for datasets where appropriate). For partner-restricted data, external users are bound by Data Sharing Agreements drafted with Imperial's Research Office, setting out the purpose of use, retention, security controls, reuse restrictions and audit rights.

6. Responsibilities

6. Responsibilities

- **Study-wide data management:** the senior PDRA (Research Fellow), under PI oversight.
- **Metadata creation:** each researcher for their own work, with the senior PDRA holding the project-level schema and ensuring cross-team consistency.
- **Data security:** Imperial ICT and Research Computing Service for infrastructure-level security; PI for project-level access decisions.
- **Quality assurance of data:** PI plus senior PDRA, with paper-level peer review providing external validation.

7. Relevant policies

7. Relevant institutional, departmental or study policies on data sharing and data security

Policy	URL or Reference
Data Management Policy & Procedures	https://www.imperial.ac.uk/media/imperial-college/research-and-innovation/research-office/public/Imperial-College-RDM-Policy.pdf
Data Security Policy	https://www.imperial.ac.uk/media/imperial-college/administration-and-support-services/secretariat/public/college-governance/charters-statutes-ordinances-regulations/policies-regulations-codes-of-practice/information-systems-security/Information-Security-Policy-v7.0.pdf
Data Sharing Policy	
Institutional Information Policy	
Other	go-fair.org/fair-principles
Other	spiral.imperial.ac.uk

8. Author and contact details

8. Author of this Data Management Plan (Name) and, if different to that of the Principal Investigator, their telephone & email contact details

Pietro Ferraro