# Offer #2024-07484

# PhD Position F/M Workflow Provenance and Its Application to Explainable and Transparent Artificial Intelligence

**Contract type :** Fixed-term contract

**Level of qualifications required :** Graduate degree or equivalent

**Other valued qualifications :** Master's degree

**Fonction :** PhD Position

## About the research centre or Inria department

The Inria Centre at Rennes University is one of Inria's eight centres and has more than thirty research teams. The Inria Centre is a major and recognized player in the field of digital sciences. It is at the heart of a rich R&D and innovation ecosystem: highly innovative PMEs, large industrial groups, competitiveness clusters, research and higher education players, laboratories of excellence, technological research institute, etc.

## Context

### Supervisory Team

- Silvina Caino-Lores, PhD (Inria, France)
- Alexandru Costan, PhD, HDR (INSA Rennes, France)
- Rafael Ferreira da Silva, PhD (Oak Ridge National Laboratory, USA)
- Ana Trisovic, PhD (Massachusetts Institute of Technology, USA)

### Location and Mobility

The thesis will be hosted by the KerData team at the Inria research center of Rennes. Rennes is the capital city of Britanny, in the western part of France. It is easy to reach thanks to the high-speed train line to Paris. Rennes is a dynamic, lively city and a major center for higher education and research: 25% of its population are students.

This thesis will include collaborations with international partners from the USA, thus research visits to and from the collaborator's teams are expected.

### The KerData team in a nutshell for candidates

- KerData is a human-sized team currently comprising 5 permanent researchers, 2 contract researchers, 1 engineer and 5 PhD students. You will work in a caring environment, offering a good work-life balance.

- KerData is leading multiple projects in top-level national and international collaborative environments such as within the Joint-Laboratory on Extreme-Scale Computing: https://jlesc.github.io. Our team has active collaboration with high-profile academic institutions all around the world (including the USA, Spain, Germany or Japan) and with industry.

- Our team strongly favors experimental research, validated by implementation and experimentation of software prototypes with real-world applications on real-world platforms incluing some of the most powerful supercomputers worldwide.

- The KerData team is committed to personalized advising and coaching, to help PhD candidates train and grow in all directions that are critical in the process of becoming successful researchers.

- Check our website for more about the KerData team herehttps://team.inria.fr/kerdata/

## Assignment

### Context and Overview

Artificial Intelligence (AI) is driving scientific discovery and economic growth in all kinds of application domains while impacting from routine daily tasks to societal-level challenges. However, research communities, industry players and social actors are expressing increasing concern about the potential ethical and practical implications of the pervasive presence of AI. Of particular concern are theexplainability of AI, or making AI's decision-making process understandable, andtransparency of AI, ensuring clarity in AI's design, data and operation. Therefore, working towards advancing explainability and transparency of AI is currently a priority, essential for responsible and trustworthy AI applications. To address these challenges, the FAIR principles (i.e., findability, accessibility, interoperability, and reuse of digital assets) have emerged as a valuable framework [WDA+16]. However, FAIRness in AI goes beyond the mere organization and sharing of data and code, encompassing the entire workflow that shapes AI models and applications.

Recent works suggest that workflow provenance (i.e., the documentation and tracking of all processes within AI development) might hold the key to supporting FAIR and Responsible AI [SAL+22, KNHJ+23]. Workflow provenance refers to capturing detailed information about all activities, processes, and transformations applied to data and code during AI development and operations.It includes information about data sources, data preprocessing, model selection, hyperparameter tuning, and evaluation metrics, among others. Capturing this provenance could provide a holistic view of the AI workflow, making it transparent and reproducible. However, a challenging aspect of working with AI workflows is that today there are no comprehensive formalisms able to capture the complexity and relationships in workflow and model provenance data [BBFM23]. Furthermore, multiple technical challenges arise when attempting to capture, store and manage the full provenance of AI workflows [MCSAGBS21, SS23], and it is still not well understood what information is valuable and how it can be leveraged in the support of AI transparency and explainability [JRG+20]

## Research Objectives

This project aims to advance the research on AI's transparency and explainability, addressing the growing concerns about ethical and practical implications of AI applications. It will investigate mechanisms
to formalize, capture, store, and manage metadata in AI-powered workflows, and will explore the relationship between model provenance, metadata, and model behavior, aiming to decipher how architectural and algorithmic characteristics impact in the model's outcome. The project is structured into three primary objectives:

1. Aim A, that focuses on the definition of ontologies and taxonomies for AI workflow provenance data from multiple angles: system (e.g., hardware, computing infrastructure, storage), platform (e.g., workflow manager, machine learning framework), model (e.g., hyperparameters, performance, architecture), and application (e.g., input and intermediate data, feedback). The outcome of Aim A is a formal and theoretical framework able to systematically capture the complexity
of the provenance metadata landscape, and facilitate a reduction of scope for the different stakeholders involved AI applications.
2. Aim B, that establishes the technical foundation to capture, store, manage and query provenance metadata at runtime during the execution of AI workflows. This includes defining data structures, algorithms, system architectures and interfaces to efficiently produce and query a detailed record of data sources, processing steps, and model configurations. The targeted main outcome of Aim B is a proof-of-concept for a large-scale provenance data management system suitable for AI workflow applications.
3. Aim C, that develops a methodology to elucidate the connections between the formalized provenance
metadata and model behavior, assessing how these elements influence model performance and interpretability. The methodology aims to evaluate the transparency and explainability in practical open-source AI models, including foundation models, in order to find links between the provenance metadata and their architectural and behavioral traits. The anticipated result of the endeavor in Aim C is a methodological framework leveraging provenance matadata taxonomies, the causal model between the studied models and their behavior, and associated statistical findings
that support transparency and explainability in practice.

# Main activities

## Envisioned Approach

To explore what AI workflow provenance metadata can tell us about AI transparency and explainability, we will build upon previous work and active research of the members in the supervisory team in the USA and France.

For Aim A, the research methodology centers on analysing the AI model life-cycle, enabling technologies and infrastructures, using our previous work on taxonomies for neural network metadata [RCLJT22] as a starting point. We described the structure of a neural network with an architectural taxonomy capturing the number, type (e.g., convolution or pooling), shape, and order of layers;

and the hyperparameters associated with each layer (e.g., kernel, stride, and padding for convolutional or pooling layers). The architectural taxonomy is independent of the data, thus allowing for comparison across datasets. We also defined a behavioral taxonomy throughout training covering the training parameters (e.g., learning rate and batch size); the criterion used for gain or loss; the method used for training and the measurement used for fitness; and the type of learning curve including, for example, designations of late-learners and never-learners. Following a similar approach, we aim to formalize similar taxonomies for other machine learning methods. This work will be complementary to our ongoing collaboration with the Workflows and Ecosystems Group from Oak Ridge National Laboratory (ORNL), in which we are exploring extensions to foundational work on provenance taxonomies with a focus on system telemetry metadata. A combination of both approaches is necessary to deliver a comprehensive formal framework suitable for developers and practitioners.

Aim B will build upon E2CLab [RCAV21], our solution for reproducible workflow execution with support for capturing provenance and monitoring metadata. Currently, E2Clab includes a provenance service that delegates system monitoring to third-party libraries in a non-intrusive way. However, E2CLab will have to be extended to provide fine-grained access to cross-layer metadata via multiple dedicated services. We hypothesize that such design will enable a more efficient deployment since multiple services can have separation of concerns in the scope of the metadata they capture (i.e., hardware, system, model and application). In addition, we expect to leverage previous work [KRCLJT22] to develop a new mechanism to capture model-specific metadata as part of this service suite. Significant efforts will be necessary to design a high-performance metadata storage middleware suitable to connect E2Clab with Flowcept, a data integration system that captures and queries workflow provenance developed by our ORNL collaborators. Finally, our preliminary results suggest algorithmic improvements will be necessary to optimize Flowcept and ensure it is not introducing overheads in the overall workflow execution.

For Aim C, we will start by analysing our collection of neural network record trails under the light of the new taxonomy from Aim A. In these previous work we amassed and annotated the life-cycle of 6,000 randomly-generated NNs across their generation, training, and validation stages [RCLJT22]. The resultant record trails, comprising both structural and learning curve data, were systematically organized in tabular text files. These record trails constitute a valuable curated collection of provenance information encompassing architecture, metadata, and performance metrics. Our collaborators from the Massachusetts Institute of Technology (MIT) are currently carrying out a large survey on opensource models that includes metadata, size and other existing metrics. In a similar approach to our previous work, we plan to generate record trails from the models provided by our collaborators, and we will enrich them with comprehensive metadata captured using the proof-of-concept from Aim B. We will apply causal inference techniques on the taxonomy-structured metadata to understand the feature strength on these data and the causal relationships between the architectural features (e.g., hyperparameters, number of layers, type of layers), behavioral features (e.g., final accuracy, accuracy curve) and other elements mapped to the taxonomies from Aim A. This approach will enable us to draw significant insights into the determinants of model interpretability, and what this can inform about transparency and explainability. We will systematically document the analysis procedure into a methodology for (i) the categorization of metadata into the aforementioned taxonomies, (ii) the extraction of key model features, and (iii) the analysis of causal relationships.

# References

[BBFM23] Elisa Bertino, Suparna Bhattacharya, Elena Ferrari, and Dejan Milojicic. Trustworthy ai and data lineage. IEEE Internet Computing, 27(6):5–6, 2023.

[JRO+20] Fariha Tasmin Jaigirdar, Carsten Rudolph, Gillian Oliver, David Watts, and Chris Bain. What information is required for explainable ai? : A provenance-based research agenda and future challenges. In 2020 IEEE 6th International Conference on Collaboration and Internet Computing (CIC), pages 177–183, 2020.

[KNHJ+23] Amruta Kale, Tin Nguyen, Frederick C Harris Jr, Chenhao Li, Jiyin Zhang, and Xiaogang Ma. Provenance documentation to enable explainable and trustworthy ai: A literature review. Data Intelligence, 5(1):139–162, 2023.

[KRCLJT22] Ariel Keller Rorabaugh, Silvina Ca´ıno-Lores, Travis Johnston, and Michela Taufer. Building high-throughput neural architecture search workflows via a decoupled fitness prediction engine. IEEE Transactions on Parallel and Distributed Systems, 33(11):2913–2926, 2022.

[MCSAGBS21] Mar¸cal Mora-Cantallops, Salvador S´anchez-Alonso, Elena Garc´ıa-Barriocanal, and Miguel-Angel Sicilia. Traceability for trustworthy ai: A review of models and tools. Big Data and Cognitive Computing, 5(2), 2021.

[RCAV21] Daniel Rosendo, Alexandru Costan, Gabriel Antoniu, and Patrick Valduriez. E2clab: Reproducible analysis of complex workflows on the edge-to-cloud continuum. In IPDPS 2021-35th IEEE International Parallel and Distributed Processing Symposium, 2021.

[RCLJT22] Ariel Keller Rorabaugh, Silvina Ca´ıno-Lores, Travis Johnston, and Michela Taufer. High frequency accuracy and loss data of random neural networks trained on image datasets. Data in Brief, 40:107780, 2022.

[SAL+22] Renan Souza, Leonardo G Azevedo, V´ıtor Louren¸co, Elton Soares, Raphael Thiago, Rafael Brandao, Daniel Civitarese, Emilio Vital Brazil, Marcio Moreno, Patrick Valduriez, et al. Workflow provenance in the lifecycle of scientific machine learning, 2022.

[SS23] Marius Schlegel and Kai-Uwe Sattler. Mlflow2prov: extracting provenance from machine learning experiments. In Proceedings of the Seventh Workshop on Data Management

for End-to-End Machine Learning, pages 1–4, 2023.
[WDA+16] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship. Scientific data, 3(1):1–9, 2016.

## Skills

**Required:**

- An excellent academic record in computer science courses
- Knowledge on distributed systems and data management systems
- Strong programming skills (Python, C/C++)
- Ability and motivation to conduct high-quality research, including publishing the results in relevant venues
- Very good communication skills in oral and written English
- Open-mindedness, strong integration skills and team spirit

**Appreciated:**

- Knowledge on machine learning and data analysis methods
- Professional experience in the areas of HPC and Big Data management

## Benefits package

- 
    - Subsidized meals
    - Partial reimbursement of public transport costs
    - Possibility of teleworking (90 days per year) and flexible organization of working hours
    - Partial payment of insurance costs

## Remuneration

Monthly gross salary amounting to 2100 euros for the first and second years and 2190 euros for the third year

## General Information

- **Theme/Domain :** Distributed Systems and middleware
  Information system (BAP E)
- **Town/city :** Rennes
- **Inria Center :** Centre Inria de l'Université de Rennes
- **Starting date :** 2024-09-01
- **Duration of contract :** 3 years
- **Deadline to apply :** 2024-06-03

## Contacts

- **Inria Team :** KERDATA
- **PhD Supervisor :**
  Caino Lores Silvina / silvina.caino-lores@inria.fr

## About Inria

Inria is the French national research institute dedicated to digital science and technology. It employs 2,600 people. Its 200 agile project teams, generally run jointly with academic partners, include more than 3,500 scientists and engineers working to meet the challenges of digital technology, often at the interface with other disciplines. The Institute also employs numerous talents in over forty different professions. 900 research support staff contribute to the preparation and development of scientific and entrepreneurial projects that have a worldwide impact.

## The keys to success

The candidate will have to show motivation, autonomy and an ability to initiate links between the research activities carried out at the INRIA center.

**Warning** : you must enter your e-mail address in order to save your application to Inria. Applications must be submitted online on the Inria website. Processing of applications sent from other channels is not guaranteed.

## Instruction to apply

Please submit online : your resume, cover letter and letters of recommendation eventually

**Defence Security :**
This position is likely to be situated in a restricted area (ZRR), as defined in Decree No. 2011-1425 relating to the protection of national scientific and technical potential (PPST).Authorisation to enter an area is granted by the director of the unit, following a favourable Ministerial decision, as defined in the decree of 3 July 2012 relating to the PPST. An unfavourable Ministerial decision in respect of a position situated in a ZRR would result in the cancellation of the appointment.

**Recruitment Policy :**
As part of its diversity policy, all Inria positions are accessible to people with disabilities.