Decentralized and Collaborative AI for Data Spaces

FUJITSU

Fujitsu Research, in cooperation with Fraunhofer ISST

Executive Summary

The interest in private AI solutions that securely leverage an organisation's unique information assets is growing, due to its promise of unlocking competitive advantages and support advanced decision-making. However, the data held by individual companies is often limited, which hinders its application to more complex business issues and the realisation of advanced predictive analytics. Data Spaces, decentralised data sharing platforms that focus on data sovereignty, offer a promising approach to tackle this challenge: Data Spaces enable secure data sharing between companies and greatly expand the potential of private AI models by integrating insights that cannot be obtained by individual companies.This white paper explains the concepts and technical directions to realise different types of private AI solutions based on Data Space principles and infrastructure. Specifically, we present technical challenges, such as data scarcity, security, and data governance, in the realisation of use cases including predictive quality control, predictive maintenance, and supply chain optimisation. We subsequently focus on three types of approaches to address these challenges: i) improving AI model training through data sharing, ii) improving the performance of AI models by leveraging knowledge from related models and data from other organisations, and iii) tackling complex business tasks through AI agent collaborations across organisational boundaries.This white paper demonstrates the potential of private AI solutions utilising Data Spaces and presents the technical initiatives that the industry and Fujitsu should pursue to realise this potential.

Table of contents

1. Introduction and Motivation

Artificial Intelligence (AI) solutions, especially Large Language Models (LLMs) of late, are gaining immensely in popularity and are becoming indispensable in virtually almost aspects of business, science, and society. Usually, Generative AI (Gen AI) models like the recent generation of flagship LLMs, have been able to acquire vast general knowledge and substantial reasoning abilities by learning from publicly available information. However, to establish a competitive advantage and support more advanced decision-making for very specific business challenges, there is an increasing need for AI solutions that leverage an organisation's proprietary, non-public information – their *private data* – and develop private AI models accordingly.

Private AI opens new doors for advanced and highly specialised solutions by learning from internal data such as customer data, operational data, and R&D data. This enables optimisation of business processes and the discovery of unique business insights that previously were impossible to achieve with AI solutions that only learned from publicly available information. At Fujitsu, we therefore believe that leveraging private AI will give companies a powerful tool to establish new competitive advantages.

However, Gen AI models usually require vast amounts of training data to be fine-tuned to a specific application - and the data held by a single company can be limited. This means that data scarcity can quickly become a bottleneck when trying to develop AI solutions for more complex business tasks or multi-dimensional challenges such as predictive analysis.

This has led Fujitsu to focus on new approaches to leverage the privately held data of different companies. By combining data held by different organisations, it is possible to gain new perspectives and insights that cannot be obtained by a single company, leading to the creation of greater value from AI models. For example, companies in the manufacturing, logistics, and retail industries could share data to optimise their supply chain management, or chemical manufacturers, automobile manufacturers, and aircraft manufacturers could share data to develop new materials.

However, when multiple companies collaborate to develop and utilise new AI solutions, it is crucial to have mechanisms in place to ensure that each company's intellectual property and other sensitive information are only shared in accordance with existing policies and agreements, and are secured. Such collaborative AI development creates the need for new frameworks that maximises the value of data integration while ensuring data security, privacy and sovereignty.

Data Spaces are attracting attention as a promising ecosystem to implement these frameworks. Data Spaces are decentralised data sharing platforms built around the concept of data sovereignty. Data sovereignty is the idea that an entity that provides data owns the rights and responsibilities related to its use. In Data Spaces, companies can control their own data whilst safely sharing it with other companies to the extent necessary. This promotes data sharing within and across industries, enabling more advanced AI development and utilisation.

While Data Spaces greatly expand the potential of private AI development, there are still technical challenges to be addressed. A recent Data Space Support Centre white paper (Data Spaces Support Centre, 2024) discusses the application of Gen AI to Data Spaces from a wide range of perspectives, including legal and technical aspects – but does not delve into the details of the architecture or technologies that could power the trend towards decentralised and private Gen AI solutions: among others, Federated Learning (FL), Retrieval-Augmented Generation (RAG), and secure computation are examples of candidate technologies that protect corporate data while enabling cooperation between companies in AI development. However, it is important to clearly specify the methods for linking these solutions with Data Space-related technologies, including how to integrate them into Data Spaces and how to determine fair incentives and compensation policies for companies participating in collaborative model training, in order to realise this vision of decentralised, collaborative AI development.

This white paper discusses key concepts and technical directions for realising private AI in Data Spaces to address these challenges. The paper is structured as follows: Chapter 2 presents three concrete use cases for decentralised AI solutions in business applications. Chapter 3 describes existing technologies and their challenges in realising these use cases, and Chapter 4 provides an overview of our approaches to addressing them. Chapter 5 explores initiatives driving the implementation of decentralised AI solutions, and Chapter 6 concludes the paper.

2. Use Case Scenarios for Collaborative AI

2.1. Predictive Quality Control

Quality Control refers to a set of processes, techniques, and systems which ensure that products or components in manufacturing meet specified standards and quality requirements. Traditionally, quality control employs statistical analysis, manual inspections, and rule-based acceptance criteria. However, with the advent of Industry 4.0, and an increasing digitisation of the shop floor that provides more production insights than ever, AI-driven approaches have revolutionised quality control by enabling faster and more accurate quality assurance.

Besides simply detecting faulty components, AI is also increasingly used for predictive quality control, i.e. the data-driven estimation of product quality based on accumulated or live process data. This enables manufacturers to implement a preventive approach rather than a reactive one by either simulating production processes or analysing real-time production insights and detecting potential quality issues before they arise.

However, data scarcity tends to be an issue for developing AI solutions aimed at predictive quality control: one company alone usually lacks sufficient data to develop a high-accuracy model for a very specific task like this. As a result, the use of collaborative AI technologies provides a promising option. Organisations with similar production lines stand to benefit from pooling their process knowledge and developing more accurate and robust tooling to improve predictive quality control and reduce spending on faulty parts. In the following chapters, we will explore the potential of leveraging Data Spaces and collaborative AI technologies to overcome data scarcity limitations and unlock valuable operational insights.

2.2. Predictive Maintenance

In the manufacturing field, a breakdown of production machinery can bring the entire production line to a halt and result in significant losses. Therefore, machines and equipment are usually maintained in a preventive fashion, and parts are replaced long before they would fail. This means that as a trade-off for uninterrupted production, replacement part lifetimes are often not exhausted efficiently, which leads to sub-optimal maintenance costs.

As a result, the concept of Predictive Maintenance (PM) has been attracting attention as a method to maintain manufacturing equipment and machines in a more resource and cost-efficient way. Traditional PM approaches have focused on developing solutions specialised in predicting equipment failure through e.g. statistical means based on historical low-dimensional data. In contrast to these traditional solutions, modern AI models are capable of considering far more complex, highly multi-dimensional data sets. This enables AI-driven PM to consider multiple facets of complex, heterogeneous manufacturing environments simultaneously and reveal potential interrelationships between variables.

However, to enable high-accuracy failure prediction using modern AI tools, it is necessary to learn wide range of failure patterns under various operating conditions. This introduces similar data scarcity issues as described above: one factory or one company alone usually does not own enough historical data to exhaustively cover all conditions and parameters required to allow for training an AI model with improved failure prediction capabilities. Because of this, AI is rarely viewed as an attractive alternative to conventional methods.

AI-powered Predictive Maintenance are therefore a prime use case for decentralised, collaborative AI solutions. Manufacturers using machines of the same supplier can pool their historical data on part failures and maintenance schedules and together develop a robust and highly accurate prediction model. To do so, they could either leverage the secure data sharing infrastructure of Data Spaces to exchange machine data for AI model training, or collaboratively leverage decentralised AI approaches.

2.3. Supply Chain Optimisation

Let's consider an example of Supply Chain Optimisation through Data Spaces, using a manufacturing scenario as an example. In this scenario, a manufacturer provides products to consumers through a chain of multiple stakeholders, including factories, logistics companies, wholesalers, and retailers. Each stakeholder performs various tasks, such as demand forecasting, production planning, inventory management, and delivery planning, and each has developed and operates their own AI model. All of these stakeholders have now joined a Data Space and are working together to optimise the entire supply chain by linking the AI models at each location.

Each retailer operates a demand forecasting AI model that has been trained using past sales data, customer data, weather data, customer reviews, and social media

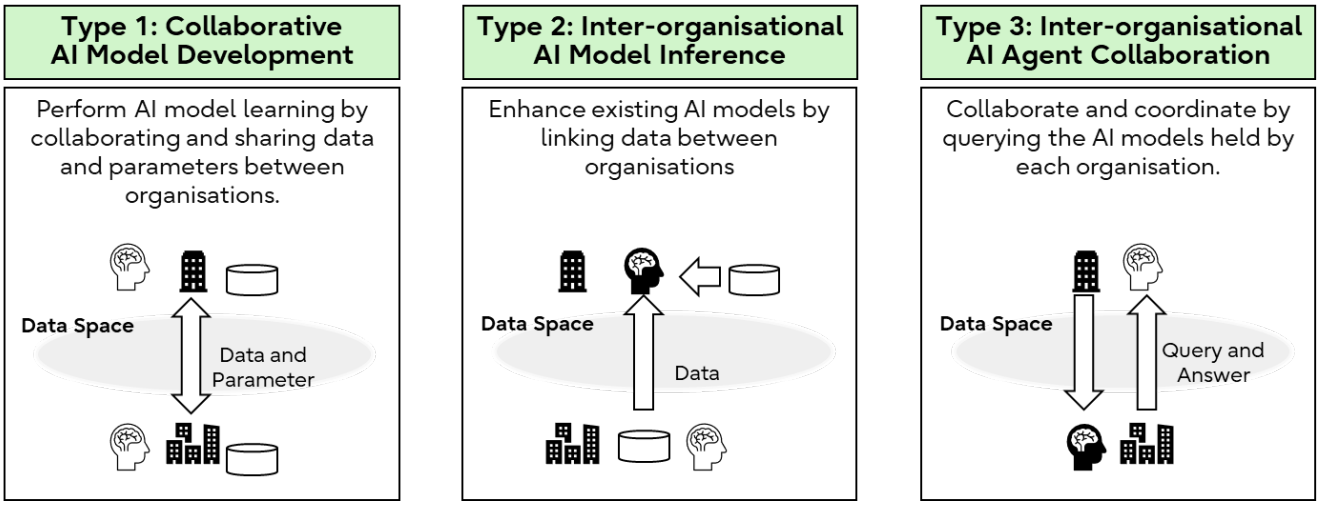

Figure 1 Three types of collaborative AI solutions

posts. By leveraging AI-generated insights, it is possible to not only predict demand but also provide product recommendations tailored to customer preferences. Wholesalers operate demand forecasting AI models that take into account not only retailers' demand forecasts but also regional event information and competitor trends. When a new product is launched the wholesaler's AI model can analyse sales data for similar products, market research data, and social media reviews to predict initial demand, enabling the retailer to secure appropriate inventory and avoid missed sales opportunities.

Sales data held by different retailers on the other hand may be difficult to share directly between stores, as they may be operated by different legal entities in different regions, such as franchises. Enabling each store to learn using its own data, and only share the results of individual learning to be aggregated in the Data Space to train the overall model, would result in the building of a higher-performance AI model. In this scenario, the data from each store is not shared with other stores, and only the knowledge of the AI model is shared, thereby protecting customer privacy while improving the performance of the AI model.

Manufacturers' production planning AI models can formulate production plans and promotions for new products based on initial demand forecasts from wholesalers, enabling smooth supply to the market. Furthermore, by sharing customer preference information generated by retailers' AI models with manufacturers' AI models, manufacturers can design and manufacture products tailored to customer needs.

This scenario showcases how multi-facetted the requirements to developing collaborative, private AI solutions can be: in some cases, simply sharing relevant data or integrating external data assets can be enough to improve AI model training and performance. In other cases, data cannot be shared directly and organisations will want to leverage decentralised AI approaches to pool their knowledge in collaborative AI solutions. And finally, in some cases organisations already developed their own, individually fine-tuned AI models. In this case – rather than developing new models - these organisations could benefit from improving their individual models by directly transferring knowledge between them, or from letting their models and AI agents communicate with each across organisational boundaries.

3. Types of Collaborative AI Solutions

In order to realise the use cases for the different private AI solutions described in the previous section, it is necessary to collaborate and share data and AI models held by different organisations. When conceptualised as an architecture, we suggest considering three different collaboration types as shown in Figure 1:

- Type 1: Collaborative AI Model Development
  Improving AI model development by integrating training data and operational knowledge from different organisations.
- Type 2: Inter-organisational AI Model Inference
  Improving AI model inference by augmenting the model's context with privately held data from different organisations.

- Type 3: Inter-organisational AI Agent Collaboration
  Tackling complex, interdependent tasks through sovereign, inter-organisational collaborations of specialised AI agents.

Type 1 is to create an AI model that aggregates the knowledge of multiple organisations by linking and sharing the data itself among organisations to accomplish a task. Type 2 is based on an AI model owned by one organisation and uses data from other organisations to supplement the knowledge of the AI model to accomplish tasks. In Type 3, each organisation already possesses a trained AI model, and accomplishes tasks by enabling a collaboration between these AI models.

This section will describe each of these three types of collaboration in more detail, and section 4 will present Fujitsu's approaches to developing the technology needed to realise business-ready solutions for each type of private AI development.

3.1. Type 1: Collaborative AI Model Development

AI solutions are now being used or explored for virtually all aspects of business. Especially Generative AI (Gen AI) solutions however are pre-trained on publicly available data, and require dedicated fine-tuning to unleash their full potential for specific applications. The more complex an application, the larger and richer needs to be the training data used to fine-tune a model. This means for a wide range of highly specialised application scenarios, organisations require more – and more diverse - training data than they can produce themselves.

Data Spaces aim to provide a secure and trustworthy ecosystem to discover and obtain these external data from other data providers. Some data however is not appropriate to be shared with others in its raw form, such as data containing sensitive operational information or privacy protected Personally Identifiable Information data. To make these types of data available for AI model training in inter-organisational collaborations, businesses need solutions that pre-process, anonymise and abstract data locally.

One approach that has been gaining attention in the past few years is the concept of Federated Learning (FL). FL is a technology that allows different contributors to develop small AI models based on their private data, and only share the parameters of these locally trained models to create a bigger, joint model. This means that with FL, private training data never leaves a contributors' premises, which improves their data sovereignty and reduces attack surfaces and the risks of leaking sensitive data. FL however also introduces a number of new challenges:

1. FL usually requires a central server to aggregate individual contributions and host the joint model – which clashes with key Data Space principles regarding to data sovereignty and decentralisation. To allow for a seamless integration of FL into Data Spaces, we thus need a fully transparent and decentralised FL solution that addresses concrete business requirements, including the sovereignty to rescind and retract previous contributions.
2. As the quality and safety of jointly developed AI solutions strongly depends on the individual model contributions, automatically detecting and removing low quality model updates and potentially malicious contributions and contributors is a key requirement for any FL solutions with intended real-world applications. In a similar vein, free riders – contributors hoping to benefit from the joint model without contributing meaningful training data – should be identified and flagged.
3. The benefits of contributing to a joint FL model might differ significantly between organisations – as might the value of their contributions. This means that in the absence of a fair framework for compensation, organisations might prefer the option to bilaterally agree on direct knowledge transfers between their models rather than contributing to a joint model. This will hold even more so when different organisations previously invested in developing their own custom models already.
4. Even when abstracting training data into model updates, sensitive information might be re-constructed or surface through prompting the joint model later on. Additional anonymisation and data security guardrails need to be implemented to equip contributors with the data government tools required to establish transparency and confidence in FL solutions.

3.2. Type 2: Inter-organisational AI Model Inference

As an alternative or extension to fine-tuning, Retrieval-Augmented Generation (RAG) has shown great potential to improve the performance of Gen AI models at inference time. RAG is a technique for retrieving information relevant to a specific query from an external knowledge base, enriching the model's context with that information, and inferring an improved response based on that information.

To leverage RAG technology for collaborative AI solutions in Data Spaces, Federated RAG (Seo, 2023) is one of the most recent approaches to link a Gen AI model to different organisations' distributed data – for example in a Data Space. Implementing Federated RAG through Data Spaces however introduces a number of challenges:

1. Due to persistent trust barriers, Data Space-connected data assets are often made available only to known partners and under pre-agreed access and usage policies. While a number of initiatives already aim to improve the process of automatically negotiating these policies for new requesters to increase the number of potential users of a given data asset, Federated RAG further intensifies this problem: Federated RAG requires automated negotiation of data usage terms between potentially thousands of data requester-provider pairs that previously had not interacted before – just for one request. This highlights the need for efficient negotiation processes to make Data Space assets available to RAG models.
2. Beyond data access agreements, enforcing data sovereignty through particular usage restrictions remains a technical challenge for Data Spaces, as there is a risk of losing control over data once it is shared with a third party. Especially for Federated RAG where information will be used by a third-party Gen AI model, it is important that data sovereignty can be guaranteed throughout the data life cycle.
3. As the performance of the Gen AI model directly depends on the quality of the retrieved information, Federated RAG also highlights the need to control the compatibility and quality of information retrieved from third-party data assets. For some applications, the currentness of information will be an issue, and new mechanisms will be needed to assess whether retrieved information is up-to-date and relevant to highly specific queries.

3.3. Type 3: Inter-organisational AI Agent Collaboration

In some cases, even pooling data from different organisations will not be enough to tackle complex, inter-dependent problems like supply chain optimisation. These problems require AI agents from different companies to interact and collaborate on a joint solution.

To make inter-organisational agent collaboration a reality, we need a trusted framework to structure and orchestrate multi-agent interactions in a way that guarantees each participant's sovereignty and enforces strict privacy requirements.

For inter-organisational AI agent interactions to be transparent and trustworthy, key principles like decentralisation, sovereignty, and data privacy and security need to be established for AI agents in the same way they have been established for data exchanges in Data Spaces before. Users need full visibility and control over agent interactions - only then will they be able to trust and leverage inter-organisational agentic AI solutions to their benefit. This means that implementing sovereign and decentralised AI agent collaboration poses a number of new challenges:

1. Agent and task discovery
   In traditional multi-agent solutions, all agents operate in the interest of the user. In order to render agents sovereign over their contribution, we need new, decentralised mechanisms that address question like: How do we discover distributed agents suitable for a given task? How do agents decide whether and how to contribute to a task? And how do agents autonomously form collaborative networks?
2. Agent communication
   When agent networks implement naïve, n-to-n communication strategies, this very quickly results in an exponential explosion in messages, problem solving duration – and costs. It also introduces risks of accidentally leaking sensitive information. This raises questions like: How do we implement efficient but decentralised agent communication? How do we prevent each agent sending messages to every other agent at every step of the collaboration process? How do we guarantee that

agent interactions do not accidentally reveal privately held information?

3. Emergent behaviour and decision making
   When agents are tasked to autonomously organise their collaboration rather than this being dictated by the user, this introduces the following challenges: How can agents collaborate efficiently while remaining sovereign and leveraging decentralised communication? How can they effectively plan together and converge on solutions?
4. Private interests and malicious contributors
   When agents are sovereign representatives of their organisations' interests, this introduces a number of key questions for organising and monitoring their collaboration: How can the agent collaboration be robust towards agents acting according to private interests and differing policies? How can malicious contributions be detected and eliminated without a central oversight mechanism?

When multiple AI agents start collaborating with each other, it is essential to verify the reliability of the AI models in order for organisations to safely interact with AI systems from other entities. This issue has already been discussed not only from a technological standpoint but also from legal systems and international standards related to AI around the world. By conducting evaluations based on established criteria, it is possible to ensure that AI models meet a certain level of quality. (European Parliament, 2024; NIST, 2023; ISO, 2023; Fairly Trained, 2024). But there are a lot of remaining issues:

1. Lack of universal standards: Each country/region has issued its own regulations and guidelines as the above activities. As a result, it is difficult to establish a globally unified certification system or standards, and certification processes become extremely complicated when AI systems collaborate across national borders (Castellvi, 2024).

2.Difficulty of updating certification criteria: It

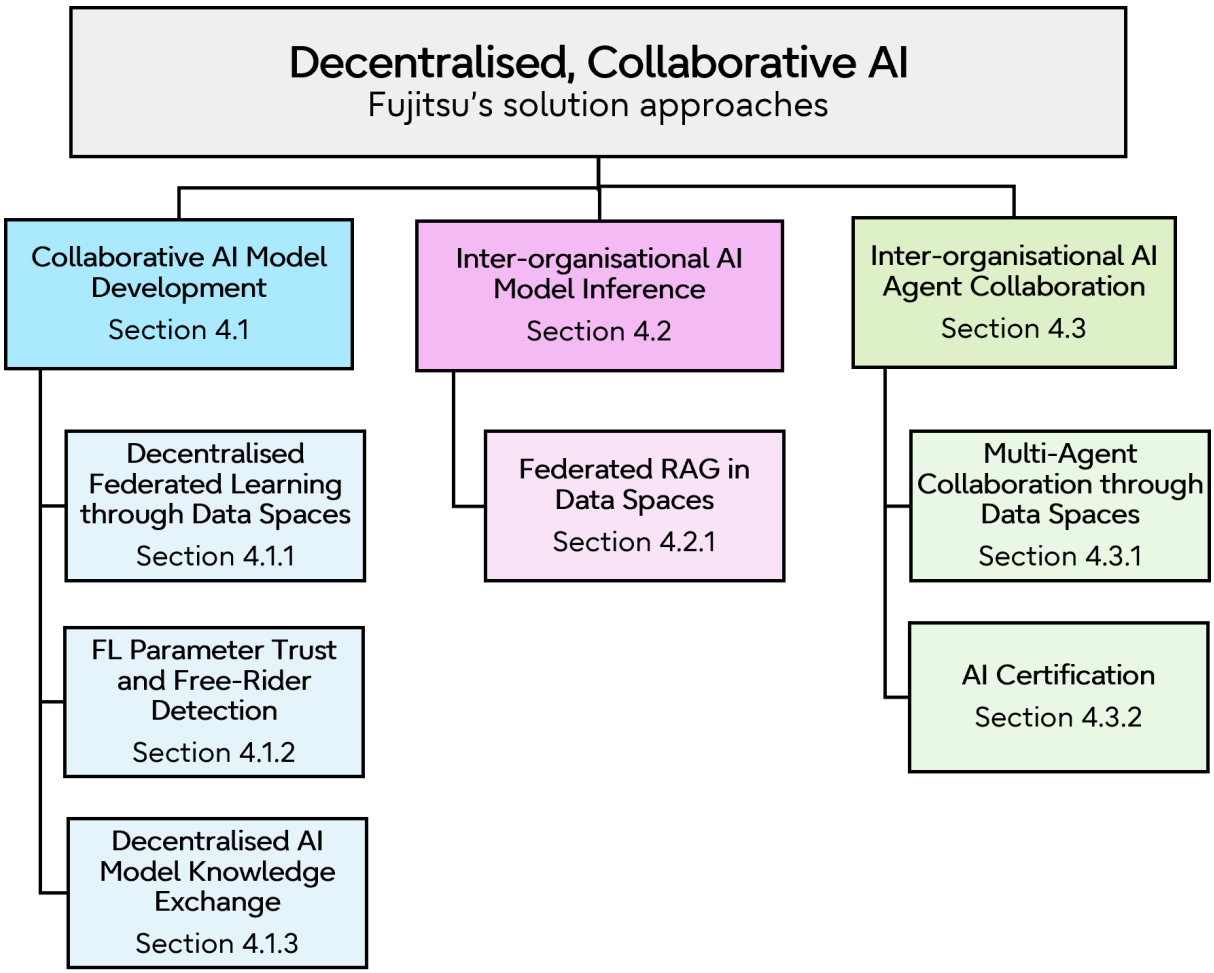

Figure 2 Overview of Fujitsu's solution approaches to the three types of AI collaboration through Data Spaces

usually takes a long time to establishing certification criteria for certification entities, and once established, subsequent changes to specifications also require significant time. This makes it difficult to update certification criteria in a timely manner to keep pace with the rapidly evolving AI landscape.

3. Update processes need automation: Many certification standards set a fixed validity period, but require frequent updates before these expiry dates because AI systems themselves are frequently updated. In such cases, if the update mechanism is not automated, the amount of certification processes becomes enormous, and the system will not work in practice.

4. Our Approach to Collaborative AI

This section introduces our efforts in resolving the above issues for each of the three different collaboration types. Figure 2 provides an overview of the initiatives described in this chapter.

4.1. Solutions for Type 1 (Collaborative AI Model Development)

4.1.1. Decentralised Federated Learning through Data Spaces

Common Federated Learning (FL) architectures stipulate a central server to aggregate all local model updates and distribute the resulting model. However, this way, a single instance decides who to provide access to the model, for what purpose and for how long. Additionally, participants are often unable to revoke their contributions, which means that they lose control over the knowledge they have provided to the joint model. On top of that, privacy remains to be a complex challenge, as past research has shown that the original data may be restored based on model updates (Zhu et al., 2019, Nasr et al., 2019), making organisations hesitant to share their valuable insights. To enable organisations to adopt FL in an industrial setting, these challenges must first be overcome.

Fujitsu and the Fraunhofer ISST have joined forces to address these challenges by developing a novel, decentralised architecture for privacy-sensitive, sovereign Federated Learning. A simple overview of this framework, implemented within the Data Space for a predictive maintenance use-case, is shown in Figure 3. Implementing this framework is not without its difficulties: decentralised FL for example turns the already complex problem of model unlearning into decentralised federated unlearning, and further raises e.g. the complexity of privacy-enhancing means like homomorphic encryption. As a result, our framework for sovereign FL will include the following two core aspects:

1. Ecosystem and Governance
   The foundation of our FL framework is built on existing Data Space principles to provide a secure and trustworthy ecosystem for exchanging and aggregating distributed model updates in a decentralised manner. This includes fair

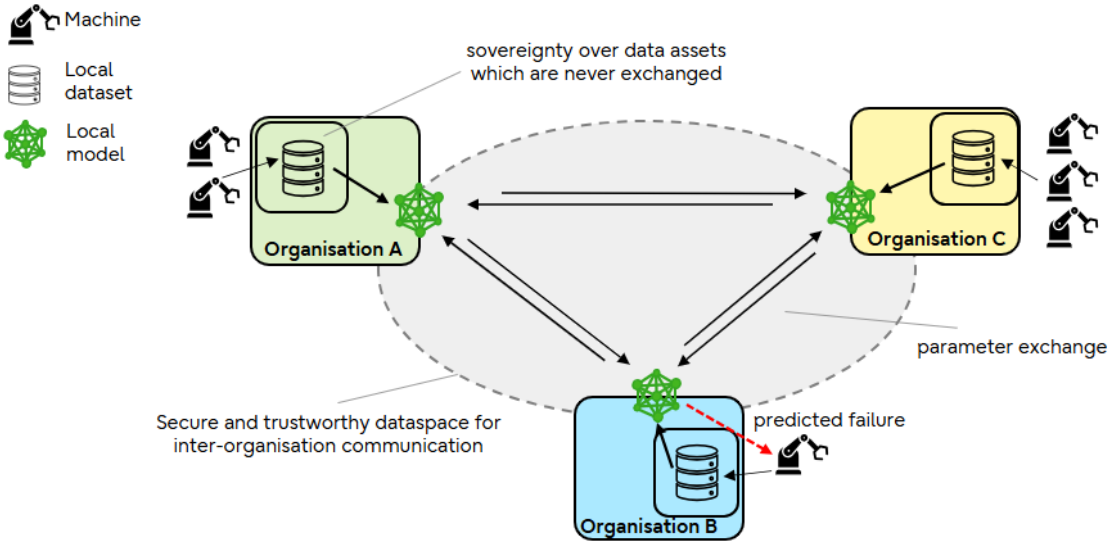

Figure 3 Decentralised Federated Learning within Data Spaces – a simple predictive maintenance use-case

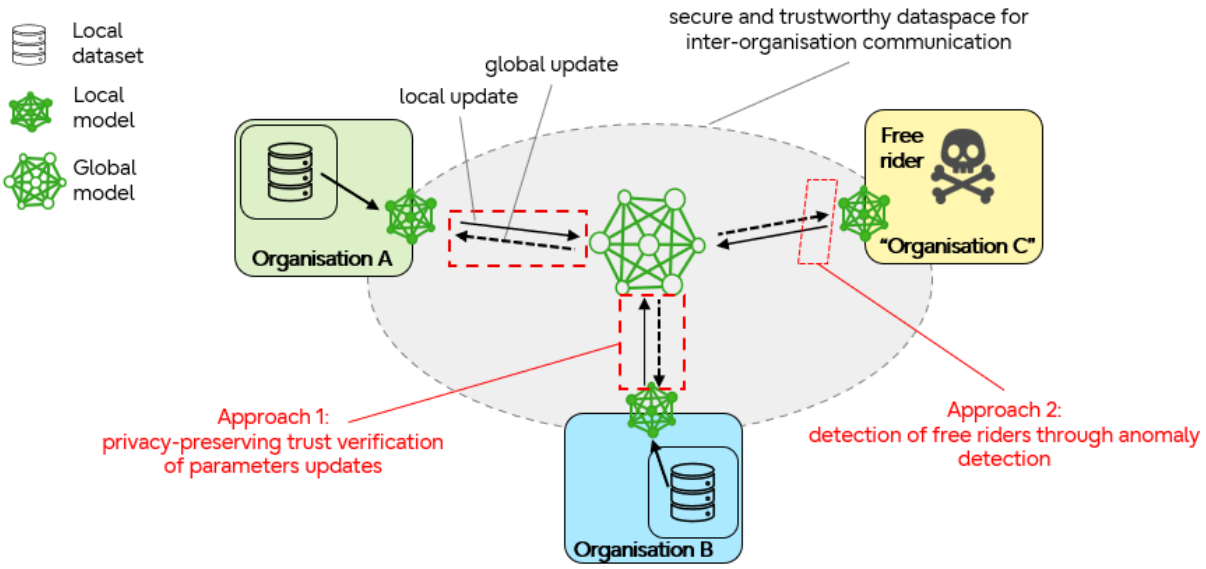

Figure 4 Authenticity verification and quality control for Federated Learning parameters

contribution incentives, interoperable and automatically enforceable contribution policies, and an extension to previous research on network topologies and model update strategies to facilitate sovereign, decentralised model training.

2. Sovereignty and Privacy
   Contributors must retain full visibility and control over their contributed knowledge before businesses will be able to leverage FL fully. To improve data privacy during model training, we are developing new mechanisms that allow for the protection of sensitive data when sharing model updates. And to implement full data sovereignty, model contributors will be provided with necessary policies and technological mechanisms to rescind and retract their model contributions.

4.1.2. FL Parameter Trust and Free-Rider Detection

In FL, AI models are developed by aggregating the parameters of locally trained contributor models. However, as parameters are simply numerical sequences, their relationship to the original learning data is unclear. For this reason, most of the commonly used data verification and quality control mechanisms are not applicable to FL. This leads to problems such as malicious contributors poisoning the model, integrating illicit back doors to the joint model, or free riders sending fake parameters to benefit from the joint model with investing in training it.

To address these issues and achieve authenticity verification and quality control for FL parameters, we are advancing research from two perspectives (Figure 4): one approach using cryptographic verification technology, and one approach using anomaly detection technology. Figure 4 illustrates the implementation of these two different, but complimentary, approaches.

In the cryptographic verification technology approach, we are researching and developing new technologies to verify the authenticity of model update parameters based on verifiable zero-knowledge proof (Chen et al., 2024; Abbaszadeh et al., 2024; Xing et al., 2023). In particular, we have proposed a method that enables verification of "learning based on raw data held by the client" without disclosing the raw data, thereby achieving both privacy and trust (Fukuoka, 2025). We aim to make this technology more practical and realise two-way proof and verification of parameter communication between clients and global models (entities that aggregate learning) to realise FL with a high level of trust.

In the anomaly detection technology approach, we are researching and developing new technologies to control parameter quality by detecting and eliminating parameters contributions that may hinder learning, such as those of free riders. As one example, we have proposed an anomaly parameter detection method that suppresses free riders (Nakamura et al., 2025). Compared to conventional methods (Li et al., 2022; Cao et al., 2020), this detection technology has the advantage

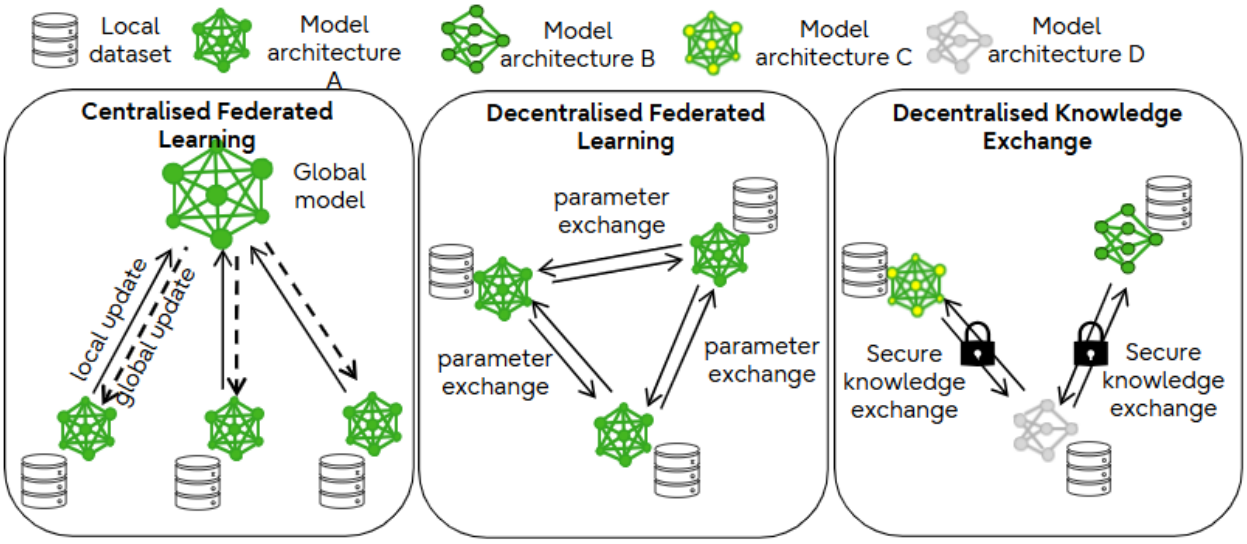

Figure 5 Overview of the information exchange mechanisms in three different frameworks: Centralised Federated Learning, Decentralised Federated Learning and Decentralised Knowledge Exchange.

of reducing the data collection cost (Mazzocca et al., 2023) on the global model side. Since the need to detect and exclude free riders is expected to increase when Federated Learning is connected with contribution incentives, we will further promote this research and work on developing more accurate and low-cost free rider detection technology.

4.1.3. Decentralised AI Model Knowledge Exchange

In some scenarios, organisations might not be interested in collaborating on a joint model. This could be the case when they value their private data higher than the expected benefits from the joint model, or when they expect that others will gain a greater benefit from the joint model than they would. In other cases, organisations might have previously developed their own models, and their chosen model architecture does not match that of a proposed joint model – which means that they would have to discard or retrain their model at potentially great cost.

To address such issues, and enable a holistic approach to AI model training in Data Spaces, whereas many collaborative LLM training scenarios as possible are considered, Fujitsu is developing a decentralised knowledge exchange framework where LLMs interact and learn from each other in a fully decentralised manner without ever exchanging model updates. A simple overview of centralised FL, decentralised FL and decentralised knowledge exchange is provided in Figure 5 to illustrate the differences between these three frameworks.

This framework by-passes potential heterogeneity bottlenecks by integrating concepts from four existing areas of technological development: i) decentralised / peer-to-peer Federated Learning; ii) parameter-efficient large-language-model (LLM) federation; iii) security-aware governance in collaborative AI; and iv) Data Space and multi-agent coordination. Furthermore, the decentralised knowledge exchange framework includes robust security guarantees to prevent malicious actors from compromising the results.

4.2. Solutions for Type 2 (Inter-organisational AI Model Inference)

4.2.1. Federated RAG in Data Spaces

In RAG, there is a technique called Federated RAG (Seo, 2023) that queries databases of multiple organisations and companies external to the user. But because data can be sensitive or highly valuable corporate data, it is necessary to establish strict contracts and access policies for the target data before making it available to a RAG system.

We are currently working on expanding Data Space infrastructure and adding data sovereignty functionality to allow for an integration of Federated RAG into Data Spaces. This functionality involves implementing local LLMs in each organisation to generate answers to RAG

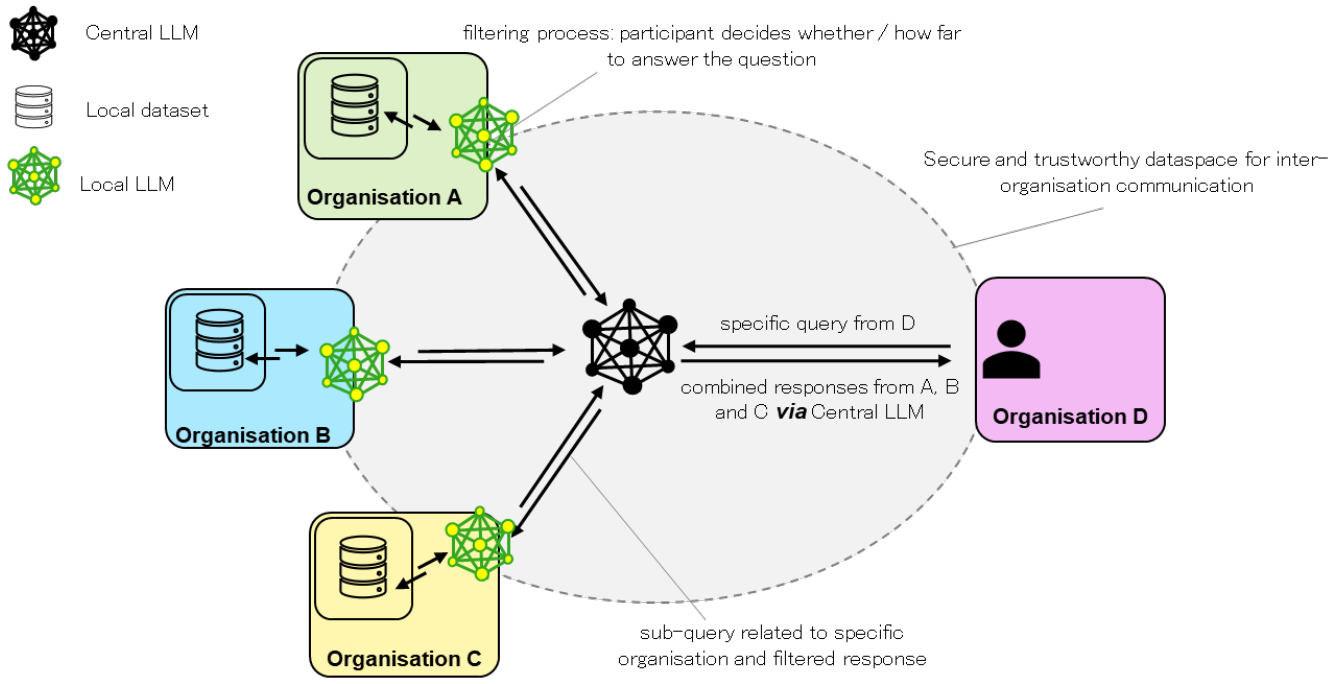

Figure 6 Federated RAG with data sovereignty functionality

questions, taking into account where the query is coming from and limiting how the answers are used by the user. It also makes it possible to balance a company's control over its data with the ease of data access facilitated by LLM utilisation.

Through these mechanisms, illustrated in Figure 6, we aim to realise a Data Space that balances the data sovereignty of organisations providing data for third-party model inference with the users' ability to maximise the utility of that data.

4.3. Solutions for Type 3 (Inter-organisational AI Agent Collaboration)

4.3.1. Multi-Agent Collaboration through Data Spaces

In conventional multi-agent collaboration, agents work together to achieve the goals specified by a single entity - typically, the user. Even when tasks are delegated to external agents hosted by other organisations, those agents usually act in the interest of the original task giver. To realise sovereign inter-organisational multi-agent collaboration, participating agents should work together to achieve a common goal whilst also each promoting the individual interests, preferences and policies of the different organisations they represent. This requires a novel method of sovereign orchestration, illustrated in Figure 7, to navigate emerging complexities and mitigate against various pitfalls.

Firstly, agents need to have the autonomy to decide whether and how to contribute to a presented task. When agents are not sovereign, usually the user or an automated controller decides what agents should contribute to solving a task – and how. But when AI agents represent individual organisations interests, these organisations' priorities and policies should govern the formation of an agent collaboration, rather than an individual requester or an external controller. To enable this, we are developing a new protocol for decentralised agent discovery and autonomous collaboration formation.

Secondly, a transparent, decentralised framework is required to host the agent collaboration. To allow for a trusted, traceable and explainable decision making process, agent interactions need to follow established protocols and leverage communication channels that uphold and enforce central sovereignty principles. Efficient autonomous decision making also requires elements like e.g. fair voting mechanisms, a record of the discussion history, and contribution provenance tracking to determine the value of individual participants' inputs. We are in the process of extending existing Data Space technologies with these features to develop a trusted, secure platform for inter-organisational AI agent collaboration.

And lastly, the orchestration of the agent collaboration process needs to be robust to potentially conflicting individual interests and the threat of untrustworthy,

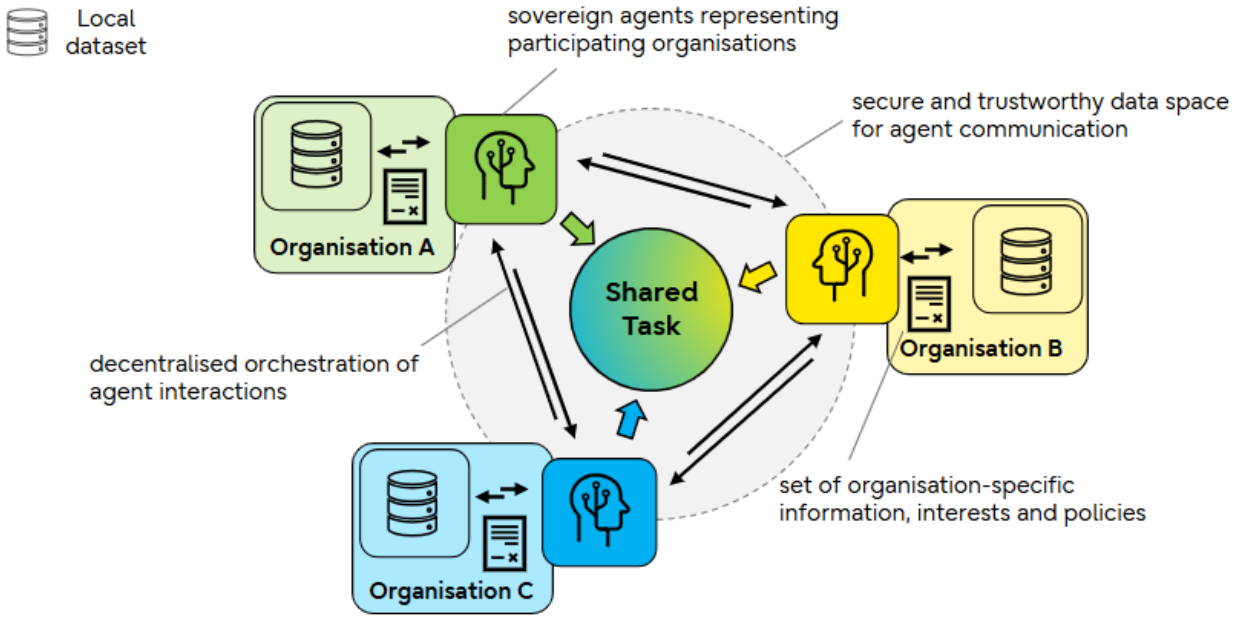

Figure 7 Sovereign multi-agent orchestration for trustworthy inter-organisational AI agent collaboration

malicious, or incompetent participants and contributions. While private interests are key to sovereign agent interactions and need to be accommodated for - and sporadic, incompetent contributions are to be expected due to the statistical nature of AI models - malicious intent and harmful contributions need to be identified and isolated immediately. To address this, we are building a novel trust framework for multi-agent collaborations to monitor, assess and evaluate the value of an agent's contributions.

4.3.2. AI certification

In order for organisations to confidently collaborate with other organisations' AI models and agents, it is important to ensure the reliability of the AI models provided as discussed in section 3.3. To this end, legal regulations, guidelines, and certification systems are being introduced. However, several challenges remain.

In particular, when aiming to automate the certification of AI systems as discussed in the third challenge in section 3.3, it is crucial to implement evidence management functions that allow not only certification bodies, but also companies across the supply chain, to automatically accumulate the data required for establishing certificates about AI systems. As a pioneering initiative, a mechanism is being developed under the name AI Bill of Materials (AI-BOM), which enables companies to maintain comprehensive lists of AI system components such as models, data, code, and infrastructure. However, at present, AI-BOM alone is not sufficient to certify AI systems.

To address this, Fujitsu is working to enhance the availability of supplementary information for reliability evaluation. We have developed a technology called Levels of Assurance for Data Trustworthiness (Data LoA) (Fujitsu Limited, 2025; Zimmer et al., 2025), which guarantees the trust level of data shared between organisations. By applying this technology to AI-BOM, it becomes possible to assess the credibility of an AI-BOM itself during AI reliability evaluations, thereby facilitating the certification process. An overview of the certification and exchange mechanism is shown in Figure 8. In addition, we are exploring what types of information and formats should be recorded to further support AI reliability assessments, contributing to the realisation of secure inter-organisational AI collaboration.

5. Trends in Related Organizations and Our Technical Proposals

The decentralised AI technologies utilizing data spaces, as described in this white paper, are being discussed in related organizations in terms of the underlying concepts and issues that form their foundation. Based on the direction of these discussions, we believe that presenting

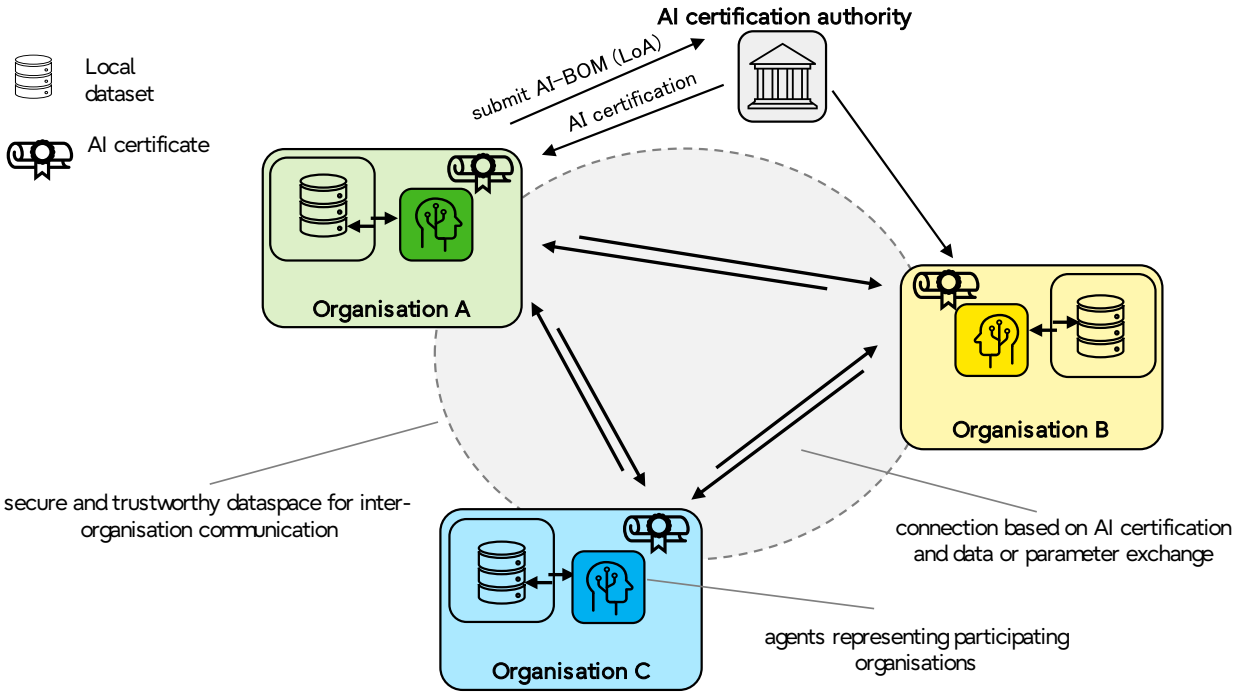

Figure 8 AI Certification and exchange mechanism

concrete technical perspectives can contribute to deepening the dialogue toward the social implementation of distributed AI. In this chapter, we organize the activities of the relevant organizations currently under consideration and show how we may contribute to them.

5.1. Council on Industrial Competitiveness – Nippon

The Japanese Council on Industrial Competitiveness – Nippon (COCN) has announced "Realisation of Socially Acceptable and Sustainable Engineering through Generative AI" (産業競争力懇談会 [COCN], 2025) as a promotion theme for the fiscal year 2025. As part of this promotion theme, COCN is investigating the additional value brought by Gen AI solutions, data infrastructure development, and ecosystem construction contributing to strengthening industrial competitiveness.

This promotion theme particularly focuses on leveraging autonomous AI agents, which are gaining attention alongside the evolution of generative AI, to achieve holistic optimization of engineering and supply chains. Key targets include the cross-organizational utilization of diverse information obtained from various sites and entities, and the development of systems in which multiple AI agents collaborate across corporate boundaries to autonomously coordinate processes and allocate resources.

Through these efforts, the promotion theme aims to enhance the resilience and efficiency of entire supply chains, enabling flexible responses to complex challenges such as disaster recovery, labour shortages, and environmental impact.

The technologies presented in Chapter 4 of this white paper are highly compatible with the themes promoted by the COCN in terms of collaboration among companies to achieve overall optimisation. The technologies in Type 1 could be used for inventory management and equipment maintenance through AI learning by linking data among companies in the supply chain and presenting incentive models for participation in AI learning; the technologies in Type 2 could be used for flexible inter-organisational information linkage using natural language; and the technologies in Type 3 could be used for linking of various AI agents owned by different organisations. We therefore will actively propose the technical directions in this white paper to this promotion theme and related activities.

5.2. Contributions to European and Japanese Data Spaces Initiatives

This white paper proposes a decentralised collaborative AI system to be realised on Data Spaces. Efforts related to Data Spaces are active in both Europe and Japan. In Europe, examples include Catena-X, Gaia-X, and IDSA,

while in Japan, examples include the Ouranos Initiative and DATA-EX. These initiatives are working to create various standards related to Data Spaces and implement them as open source software (OSS).

The technologies proposed in this white paper are designed for use on Data Spaces. Therefore, by contributing to the relevant data space standards and their OSS implementations, we aim to make it easier for stakeholders involved in Data Spaces to access and use these technologies.

6. Conclusion

This white paper discusses the concepts and technical directions for realising private AI in Data Spaces. AI that leverages organisation-specific data is essential for companies to establish competitive advantages and make advanced decisions. However, there are limits to the data that a single company can hold, making data sharing between companies important. In response, Data Spaces promote collaboration between companies as a secure data sharing platform centred on data sovereignty, enabling more advanced AI development and utilisation. This white paper introduced the challenges of realising private AI in Data Spaces and Fujitsu's efforts to address them. Going forwards, we plan to apply the technologies presented in this white paper in various settings, including Japan and Europe.

Fujitsu will continue to advance technological development and proof-of-concept experiments to realise private AI through Data Spaces, and support companies in making data-driven decisions and unlocking new business value.

References

1. Abbaszadeh, K., et al. (2024). Zero-knowledge proofs of training for deep neural networks. *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*.
2. Cao, X., et al. (2020). Fltrust: Byzantine-robust federated learning via trust bootstrapping. *arXiv preprint*, arXiv:2012.13995. https://arxiv.org/abs/2012.13995
3. Chen, B.-J., et al. (2024). Zkml: An optimizing system for ML inference in zero-knowledge proofs. *Proceedings of the Nineteenth European Conference on Computer Systems*.
4. 産業競争力懇談会（COCN）. (2025). 生成 AI による社会受容性のあるサステナブルなエンジニアリングの実現. 産業競争力懇談会. http://www.cocn.jp/report/895c2b77b0ddc05500e784e0924a7c3e27567987.pdf
5. Data Spaces Support Centre. (2024). The new “Generative AI and Data Spaces" *white paper of the Strategic Stakeholder Forum is now available*. Data Spaces Support Centre. https://dssc.eu/space/News/blog/380600324/
6. European Parliament. (2024). *EU Artificial Intelligence Act*. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689
7. Fairly Trained. (2024). Fairly Trained launches certification for generative AI models that respect creators’ rights. https://www.fairlytrained.org/blog/fairly-trained-launches-certification-for-generative-ai-models-that-respect-creators-rights
8. Fujitsu Limited. (2025). White paper on a framework for ensuring data trustworthiness published with Fraunhofer ISST. https://www.fujitsu.com/global/about/research/article/202503-data-loa.html
9. 福岡, 尊. (2025). 検証可能な連合学習方式の一提案. 2025 年 暗号と情報セキュリティシンポジウム (SCIS2025）.
10. ISO. (2023). ISO/IEC 42001:2023. https://www.iso.org/standard/42001
11. Li, Y., et al. (2022). An effective federated learning verification strategy and its applications for fault diagnosis in industrial IoT systems. *IEEE Internet of Things Journal*, 9(18), 16835–16849. https://doi.org/10.1109/JIOT.2022.3181162
12. 中村, 元紀, 他. (2025). 信頼度を加味した連合学習におけるモデルパラメータ評価手法の一提案. DEIM2025 第 17 回データ工学と情報マネジメントに関するフォーラム (第 23 回日本データベース学会年次大会）.
13. Nasr, M., Shokri, R., & Houmansadr, A. (2019). Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP)* (pp. 739–753). IEEE. https://doi.org/10.1109/SP.2019.00065
14. Mazzocca, C., et al. (2023). TruFLaaS:

Trustworthy federated learning as a service. *IEEE Internet of Things Journal*, 10(24), 21266–21281. https://doi.org/10.1109/JIOT.2023.3322617

15. NIST. (2023). AI Risk Management Framework. https://www.nist.gov/itl/ai-risk-management-framework
16. Seo, J. (2023). FedRAG: Federated Retrieval Augmented Generation. https://doi.org/10.13140/RG.2.2.20012.78727
17. Xing, Z., et al. (2023). Zero-knowledge proof meets machine learning in verifiability: A survey. *arXiv preprint*, arXiv:2310.14848. https://arxiv.org/abs/2310.14848
18. Zhu, L., Liu, Z., & Han, S. (2019). Deep leakage from gradients. In *Advances in Neural Information Processing Systems*, 32, Article 1323, 14774–14784. Curran Associates Inc.
19. Zimmer, F., Haber, J., Kaneko, M., & Takeuchi, T. (2025). Towards levels of assurance for data trustworthiness: A novel framework to promote trust in inter-organisational data sharing. In BPMS2 2025, RCIS2025.
20. Zimmer, F., Haber, J., & Kaneko, M. (2025). Enhancing Trust in Inter-Organisational Data Sharing: Levels of Assurance for Data Trustworthiness. In DATA2025.

Inquiries regarding this matter

Data & Security Research Laboratory,
Fujitsu Research,
Fujitsu Limited.
Email: fj-aifordataspaces@dl.jp.fujitsu.com

