Decentralized and Collaborative AI for Data Spaces

FUJITSU

Fujitsu Research, in cooperation with Fraunhofer ISST

Executive Summary

競争優位性の確立と高度な意思決定を支援するため、組織独自の情報を活用したプライベート AI の重要性が高まっています。しかし、単独の企業が保有するデータには限りがあり、より広範なビジネス課題の解決や高度な予測分析の実現を阻む要因となっています。この課題に対し、データ主権を軸とした非中央集権型のデータ共有基盤であるデータスペースは有望なアプローチです。データスペースは企業間での安全なデータ連携を可能にし、単独企業では得られない知見を引き出すことで、プライベート AI の可能性を大きく広げます。本ホワイトペーパーでは、データスペースにおけるプライベート AI 実現に向けた概念と技術の方向性を解説します。具体的には、予測品質管理、予知保全、サプライチェーン最適化といったユースケースを想定し、その実現におけるデータ不足、セキュリティ、データガバナンスなどの技術課題を提示します。さらに、これらの課題に対し、データ連携による AI 学習、学習済 AI モデルと複数組織データの連携、組織間の学習済 AI モデル協調という 3 つの類型に着目し、技術的解決に向けたアプローチを紹介します。本ホワイトペーパーは、データスペースを活用したプライベート AI の可能性を示すとともに、その実現に向けて業界および富士通が目指すべき技術的取り組みを提示します。

目次

1. 目的、背景

今日、人工知能（AI）技術、とりわけ生成 AI は、ビジネス、科学、社会のあらゆる側面において、その存在感を増し、不可欠なものとなりつつあります。通常、生成 AI は、Web などの公開情報を学習することで、汎用的な知識を獲得し、様々なタスクをこなせるようになりました。しかし、競争優位性を確立し、より高度な意思決定を支援するためには、組織固有の非公開情報、いわゆる「プライベートデータ」を活用した AI、すなわちプライベート AI の必要性が高まっています。

プライベート AI は、企業の内部データ（顧客データ、業務データ、研究開発データなど）を学習することで、企業活動への高度な活用という新たな扉を開きました。プライベート AI はこれまで公開情報のみを学習した AI では難しかった、企業内の特定の業務プロセス最適化や、独自のビジネスインサイト発見を可能にします。我々はプライベート AI の活用が、企業が競争優位性を確立するための強力な武器になると考えています。

しかし、生成 AI モデルを作るためには、膨大な量の学習データが必要になります。通常単一企業が持つデータには限りがあるため、より広範なビジネス課題を解決したり、より高度な予測分析を行ったりするためには、データ量の不足がボトルネックとなる可能性があります。

そこで注目されるのが、業界内、あるいは業界を超えて複数の企業が持つデータを活用するというアプローチです。異なる企業が保有するデータを組み合わせることで、単一企業では得られない新たな視点や知見が得られ、より高い価値創造ができる可能性があります。例えば、サプライチェーン全体を最適化するために、製造業、物流業、小売業の各社がデータを連携したり、新素材開発のために、化学メーカー、自動車メーカー、航空機メーカーがデータを共有したりといった活用が考えられます。

しかし、複数の企業が連携して AI を開発・活用する場合、各企業が保有する知的財産やトレードシークレット（営業秘密）を安全に共有しながら学習を進めるための仕組みが不可欠となります。データの機密性やセキュリティ、主権を確保しつつ、データ連携による価値を最大化するための新たなアーキテクチャと、それを実現するための技術が求められています。

この課題を解決する有望なアプローチとして、データスペースが注目されています。データスペースは、データ主権（Data Sovereignty）を軸として構築された、非中央集権型のデータ共有基盤です。データ主権とは、データを提供する主体が、そのデータの利用に関する権利と責任を持つという考え方です。データスペースでは、各企業は自社のデータをコントロールしながら、必要な範囲でデータを他の企業と安全に共有することができます。これにより、業界内や業界間でのデータ連携を促進し、より高度な AI 開発・活用を可能にします。

データスペースは、プライベート AI の可能性を大きく広げる一方で、現状では技術的な課題が残されています。DSSC White Paper (Data Spaces Support Centre, 2024)においては、データスペースへの生成 AI の適用に関して法的側面、技術的側面など広範に議論されていますが、非中央集権型アーキテクチャやプライベート AI を支援するような技術の詳細について踏み込んだ議論には至っていません。たとえば、連合学習や検索拡張生成(Retrieval Augmented Generation, RAG)、秘密計算などは企業のデータを守りながら企業間で協力して AI の学習を行うために使える技術の候補ですが、非中央集権型の協調 AI 開発を実現するために、これらの技術をデータスペースとどのように組み合わせればよいか、学習に参加する企業のインセンティブや対価配分をどうすればいいかなど、データスペース関連技術との具体的な連携手法を明確にする必要があります。

本ホワイトペーパーでは、これらの課題を解決するために、データスペースにおけるプライベート AI の実現に向けた概念と技術の方向性について議論します。本ホワイトペーパーの以降の構成は次の通りです。2 章では本ホワイトペーパーで想定するユースケースを挙げます。3 章ではユースケースを実現する際における既存の技術及びその課題を示します。4 章ではこれら課題に対する我々のアプローチを述べます。5 章では今後想定される適用先について述べます。最後に 6 章で全体をまとめます。

2. 想定されるユースケース

2.1. 予測品質管理

品質管理とは、製造工程における製品や部品が指定された基準や品質要件を満たすことを確保するためのプロセス、技術、およびシステムの集合体です。伝統的に、品質管理では統計分析、手動検査、ルールベースの受け入れ基準が採用されてきました。しかし、Industry 4.0 の台頭と製造現場のデジタル化が進み、これまで以上に生産に関する洞察が得られるようになったことで、AI を活用したアプローチが品質管理を革命的に変革し、より迅速かつ正確な品質保証を実現しています。

AI は、単に不良部品を検出するだけでなく、蓄積されたデータやリアルタイムの工程データに基づく製品品質の予測的な評価（予測品質管理）にも活用されてきています。これにより、製造企業は生産工程のシミュレーションやリアルタイムの生産洞察の分析を通じて、問題が発生する前に潜在的な品質課題を検出することで、事後的な対応ではなく予防的なアプローチを実施できるようになります。

しかし、予測品質管理向けの AI ソリューション開発において、データ不足は課題となっています。特定のタスク（例えば、非常に専門的な品質管理）向けに高精度なモデルを開発するには、通常、1 社だけでは十分なデータが不足します。そこで、複数企業間での連携による協業型 AI 技術の活用が有望な選択肢となります。これにより、同様の生産ラインを持つ各組織が、プロセス情報を共有し、AI を活用したより正確で堅牢な品質管理ツールを開発することで、予測品質管理の精度向上と不良部品のコスト削減を実現できます。この後の章では、データスペースや協業型 AI 技術を活用することで、データ不足の制約を克服し、より複雑な組織間の共通タスクを実行する方法について説明します。

2.2. 予知保全

製造業において、生産機械の故障は生産ライン全体を停止させ、重大な損失を引き起こす可能性があります。そのため、機械や設備は予防的なメンテナンスが行われ、部品は故障する前に交換されます。これは、生産の連続性を維持する代償として、交換部品の寿命が効率的に消費されずメンテナンスコストが最適化されないというトレードオフが生じます。

このため、製造設備や機械をより効率的に維持する手法として、予知保全が注目されています。この手法の実現のため、生産設備の稼働率を最大化するための故障予測を行う AI 技術に多くの研究が行われています。従来の予知保全は、限られた履歴データに基づく統計的手法などにより設備の故障を予測するものでしたが、現代の AI モデルはより複雑で高次元なデータセットを活用可能です。これにより、AI 駆動型予知保全は複雑で多様な製造環境の多面的な要素を分析し、変数間の構造的な依存関係を明らかにできます。

しかし、AI を用いた高精度な故障予測を実現するには、多様な運転条件下での広範な故障パターンを学習する必要があります。これには、上述と同様のデータ不足の問題が伴います。1 つの工場や 1 つの企業だけでは、AI モデルを学習するために必要なすべての条件とパラメータを網羅する十分な履歴データを持っていないことが一般的です。このため、単純な AI 活用は従来の方法の有望な代替手段として見なされない傾向があります。

この課題に対処するため、AI を活用した予知保全は、本ホワイトペーパーで論じる協業型 AI 技術のもう 1 つの主要な活用事例となります。同じサプライヤーの機械を使用するメーカーは、部品の故障データやメンテナンススケジュールを共有し、共同で堅牢で高精度な予測モデルを開発できます。そのためには、データスペースを活用して AI モデル学習のための機械データを交換したり、分散型 AI アプローチを共同で活用したりすることができるでしょう。

2.3. サプライチェーン最適化

製造業を例に、データスペースを活用したサプライチェーン最適化の事例を考えてみましょう。メーカーは、複数の工場、物流業者、卸売業者、小売業者などのステイクホルダーを通じて、商品を消費者に届けています。それぞれのステイクホルダーは、需要予測、生産計画、在庫管理、配送計画など、様々な業務が行われており、それぞれが独自の AI モデルを開発・運用しています。メーカーはデータスペースを構築し、各拠点の AI モデルを連携させることで、サプライチェーン全体の最適化を目指しています。

各小売業者は、過去の販売データ、顧客データ、気象データ、顧客レビューや SNS の投稿などを学習した需要予測 AI モデルを運用しています。AI による生成を活用することで、単に需要を予測するだけでなく、顧客の嗜好に合わせた商品リコメンデーションも行うことができます。卸売業者では、小売業者の需要予測に加え、地域ごとのイベント情報や競合店の動向などを考慮した需要予測 AI モデルを運用しています。新製品の発売時には、卸売業者の AI モデルは、類似製品の販売データ、市場調査データ、SNS の口コミ情報などを分析し、初期需要を予測することで、小売業者が適切な在庫を確保し、販売機会を逃さないよう支援することができます。

一方、小売業者が持つ販売データは、フランチャイズなど地域ごとに異なる法人が運営している場合もあり、店舗間で直接共有することは難しい場合があります。そこで、各店舗がそれぞれ自身の持つデータを使って学習を行い、個別に学習させた結果だけをデータスペース上で集約し全体モデルの学習を行うことで、より性能の高い AI モデルを構築します。各店舗のデータは互いにに共有されることはなく、AI モデルの知識だけが共有されるため、顧客のプライバシーを保護しながら、AI モデルの性能を向上させることができます。

メーカーの生産計画 AI モデルは、卸売業者からの初期需要予測に基づき、新製品の生産計画やプロモーションを立案し、市場への供給をスムーズに行うことができます。また、小売業者の AI が生成した顧客の嗜好情報を製造業者の AI に共有することで、製造業者は顧客のニーズに合わせてパーソナライズされた製品を設計・製造することもできるようになります。

このシナリオは、協業型プライベート AI ソリューションの開発要件が多岐にわたることを示しています：一部のケースでは、関連するデータを共有したり外部データ資産を統合したりするだけで、AI モデルのトレーニングと性能を向上させることができます。他のケースでは、データを直接共有できないため、組織は分散型 AI アプローチを活用し、知識を共有する方法を模索する必要があります。最後に、一部の組織では既に独自の AI モデルを開発しています。このような場合、組織はそれら間で知識移転を行うことでモデルを改善した

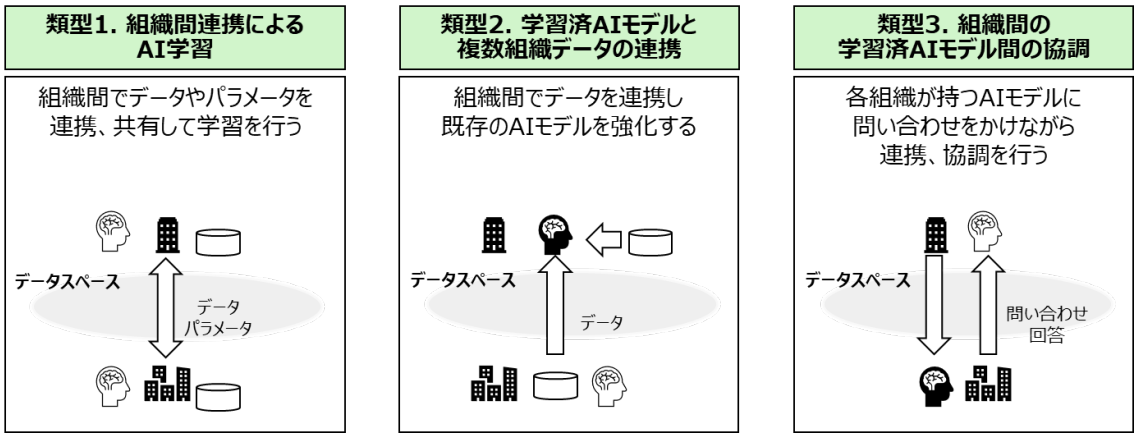

図 1 データおよび AI の組織間連携の 3 つの類型

り、組織の境界を越えてモデルと AI エージェントを相互作用させたりすることで、より大きな利益を得ることができます。

3. データおよび AI の組織間連携の類型

前章で説明したユースケースの実現にあたっては、各組織で保有するデータや AI モデルを必要に応じて連携、共有して問題解決を行っていく必要があります。

アーキテクチャとしてみた場合、この連携の方法には以下の図 1 で表されるような 3 種類の類型を考えることができます。

- 類型 1. 組織間連携による AI 学習
- 類型 2. 学習済 AI モデルと複数組織データの連携
- 類型 3. 組織間の学習済 AI モデル間の協調

類型 1 はデータそのものを組織間で連携、共有することで複数組織の知見を集約した AI モデルを作成し、タスクを遂行します。類型 2 はある組織が持つ AI モデルをベースに、他の組織のデータを利用して AI モデルの知識を補完することでタスクを遂行します。類型 3 は各組織が学習済みの AI モデルをすでに保有しており、AI モデル間で連携を行うことでタスクを遂行します。

本章では各類型について詳しく説明し、4 章にて課題解決に向けた技術について我々の取り組みを示します。

3.1. 類型 1. 組織間連携による AI 学習

AI ソリューションは現在、あらゆるビジネス領域で活用、検討されています。特に生成 AI ソリューションは、公開されているデータで事前学習されており、特定の用途においてその真価を発揮するには、専用のファインチューニングが必要です。アプリケーションが複雑になるほど、モデルをファインチューニングするために必要な学習データは、より大規模かつ多様である必要があります。つまり、専門性の高い多くのユースケースにおいて、組織は自社だけでは十分な量と多様性のあるデータを用意できないことが多いのです。

データスペースは、こうした外部データを他のデータ提供者から安全かつ信頼できる形で発見・取得するためのエコシステムを提供することを目的としています。しかし、運用上の機密情報や個人を特定できる情報など、元のままでは共有に適さないデータも存在します。こうしたデータを組織間での AI モデル学習に活用するには、データをローカルで前処理・匿名化・抽象化するソリューションが必要です。

近年注目を集めているアプローチの一つが連合学習です。連合学習は、各参加者が自分のプライベートデータを使ってローカルで小規模な AI モデルを学習し、そのパラメータのみを共有して、より大きな共同モデルを構築する技術です。これにより、学習データが外部に出ることなく、データ主権を維持しつつ、攻撃リスクや機密情報漏洩のリスクを低減できます。ただし、連合学習にはいくつかの新たな課題もあります：

1. 中央サーバーの必要性：連合学習では通常、個々の貢献を集約し、共同モデルをホストする中央サーバーが必要ですが、これはデータスペースの「データ主権」や「非中央集権性」といった原則と矛盾します。連合学習をデータスペースにシームレスに統合するには、過去の貢献の取り消し権利などを含む、透明性が高く分散型の連合学習ソリューションが必要です。
2. 品質と安全性の確保：共同開発された AI ソリューションの品質と安全性は、各参加者のモデルからの入力情報に大きく依存します。低品質なモデル更新や悪意ある貢献・貢献者を自動的に検出・排除する機能が、現実世界での連合学習活用には不可欠です。同様に、意味のあるデータを提供せずに利益だけを得ようとする「フリーライダー」も特定・対処する必要があります。
3. 貢献の価値と報酬の不均衡：共同モデルへの貢献による利益やその価値は、組織ごとに大きく異なる可能性があ

ります。公平な報酬の枠組みがなければ、組織は共同モデルへの貢献よりも、個別に知識を交換する方を選ぶかもしれません。特に、すでに独自のカスタムモデルに投資している場合はその傾向が強まります。

4. モデル更新からの情報漏洩リスク：共同モデルへの貢献のために学習データを抽象化しても、後に共同モデルに問い合わせを行うことで、機密情報が再構築されたり露出したりする可能性があります。これを防ぐためには、追加の匿名化処理やセキュリティ対策が必要であり、連合学習ソリューションに対する透明性と信頼性を確保するためのデータガバナンスツールが求められます。

3.2. 類型 2. 学習済 AI モデルと複数組織データの連携

前節で述べたファインチューニングのほかに生成 AI を強化する方法として、RAG があります。RAG は、生成 AI が受け取ったユーザーのリクエスト内容から、必要な情報を連携先のデータベースから取得し、その情報を基に推論する技術です。

データスペースにおける協調的な AI ソリューションに RAG 技術を活用するために、Federated RAG （Seo, 2023）は最新アプローチの一つです。これにより、生成 AI モデルを複数の組織に分散されたデータ、たとえばデータスペース内のデータに接続することができます。

しかしながら、データスペースを通じて Federated RAG を実装するには、いくつかの課題が存在します：

1. データスペースでは、データは通常、事前合意された特定の相手に、アクセスおよび利用ポリシーに従って提供されます。これまでの取り組みは、自動交渉プロセスを改善することで、データ提供者がデータ利用者を増やすことを目的としています。しかし、連合 RAG では、これまで相互にやり取りしたことがない数千のデータ利用者とデータ提供者の間で、単一の要求に対するデータ利用条件の自動交渉が必須となることから、この問題をより難しくします。そのため、RAG モデルにデータ資産を提供するための効率的な交渉プロセスの必要性が浮き彫りになっています。
2. データ主権を特定の利用制限を通じて強制することは、データスペースにおける技術的な課題です。データが第三者と共有されると、データ制御を失うリスクがあるためです。特に連合 RAG では、データが第三者の生成 AI モデルによって利用されるため、データ主権がデータライフサイクル全体で保証されることが重要です。
3. 生成 AI モデルの性能は取得した情報の品質に直接依存するため、フェデレーテッド RAG は第三者から取得するデータの互換性と品質を管理する必要性を浮き彫りにしています。一部のアプリケーションでは情報の最新性が問題となり、取得した情報が最新かつ関連性があるかどうかを評価するための新たなメカニズムが求められるでしょう。

3.3. 類型 3. 組織間の学習済 AI モデル間の協調

異なる組織間でデータを連携しても、サプライチェーン最適化のような複雑で相互依存性の高い問題に対処するには不十分となる場合があります。これらの問題には、異なる企業の AI エージェントが相互に連携し、共同で解決策を導き出す必要があります。

組織間でのエージェント協力を実現するためには、各参加者の主権を保証し、厳格なプライバシー要件を遵守しながら、マルチエージェント相互作用を構造化・調整する信頼できるフレームワークが必要です。

データスペースは、分散化、主権、データプライバシーとセキュリティといった主要な原則に従って組織間データ交換を実現します。組織間 AI エージェントの相互作用が透明で信頼できるものとするためには、同じ原則を AI エージェントにおいても確立する必要があります。ユーザーはエージェントの相互作用に対して完全な可視性と制御を保持する必要があります。そうすることで初めて、ユーザーは組織間エージェント型 AI ソリューションを信頼し、活用できるようになります。つまり主権的で分散化された AI エージェントの協業を実現することは、以下のような数多くの新たな課題をもたらすことを意味します：

1. エージェントとタスクの発見
   従来のマルチエージェントソリューションでは、すべてのエージェントは単一のユーザーの利益のために動作します。複数のエージェントが自身の貢献に対して自律性を確保するためには、次のような質問に対応する新しい分散型メカニズムが必要です：与えられたタスクに適した分散エージェントをどのように発見するか？エージェントは、与えられたタスクへの貢献の有無や方法をどのように決定するか？エージェントは自律的に協働ネットワークをどのように形成するか？
2. エージェント間の通信
   エージェントネットワークが単純な n 対 n の通信戦略を採用すると、メッセージの爆発的増加、問題解決時間の延長、コストの増加などが発生します。また、機密情報の漏洩リスクも生じます。これらは以下の課題を提起します：効率的で分散化されたエージェント間の通信をどのように実装するか？各エージェントが他のすべてのエージェントにメッセージを送信するのを防ぎ、メッセージの爆発的増加と実行コストの急増を回避するにはどうすればよいか？エージェント間の相互作用が、意図せず秘密情報を漏洩しないように保証するにはどうすればよいか？

3. 創発的行動と意思決定
   エージェントがユーザーによって指示されるのではなく、自律的に協業を組織化するようタスクが与えられる場合、以下の課題が生じます：エージェントが分散型通信を行いつつ、自律性を維持しながら効率的に協業する方法は？どのようにして効果的に計画を立て、解決策に収束させるか？
4. 私的利益と悪意ある貢献者
   エージェントが組織の利益を自律的に代表する場合、その協業の組織化と監視において以下の重要な質問が生じます：エージェントの協働を、私的利益や異なるポリシーに従って行動するエージェントに対して頑強にするにはどうすればよいでしょうか？中央監視メカニズムなしで悪意ある貢献を検出・排除するにはどうすればよいでしょうか？

複数の AI エージェントが上記のように協調して動作するとき、安心して各組織が他組織の AI と連携するためには、提供される AI モデルの信頼性を確認する必要があります。エージェントの中には、悪意のある行動に関与するものが紛れ込んでいる可能性もあるので、セキュリティ面でも重要な確認となります。この問題は、技術的な観点だけでなく、世界中の AI に関連する法制度や国際標準の観点からもすでに議論されており、定められた基準に基づき審査することで、AI モデルが一定の品質を満たすことを担保できます（European Parliament, 2024; NIST, 2023; ISO, 2023; Fairly Trained, 2024）。しかし、多くの課題が残されています：

1. 統一的な基準の欠如：各国・地域は上記の活動と同様に、独自の規制やガイドラインを策定しています。その結果、グローバルに統一された認証システムや基準を確立することが困難であり、AI システムが国境を越えて連携する場合、認証プロセスが非常に複雑になります(Castellvi, 2024)。
2. 認証基準の更新の難しさ：認証機関の基準を策定するには通常長い時間がかかり、一度策定されると、その後の仕様変更にも多くの時間を要します。これにより、急速に進化する AI の状況に即応して認証基準をタイムリーに更新することが困難になります。
3. 更新プロセスの自動化の必要性：多くの認証基準には有効期限が設定されていますが、AI システム自体が頻繁に更新されるため、有効期限を待たずに認証を更新する必要があるケースが多くなります。このような場合、更新メカニズムが自動化されていなければ、認証作業の量が膨大になり、実際には運用が困難になります。

4. AI 連携に向けた我々の取り組み

本章では、前章で示した類型ごとの課題に対する我々の取り組みを紹介します。図 2 に本章で説明する取り組みの概観を示します。

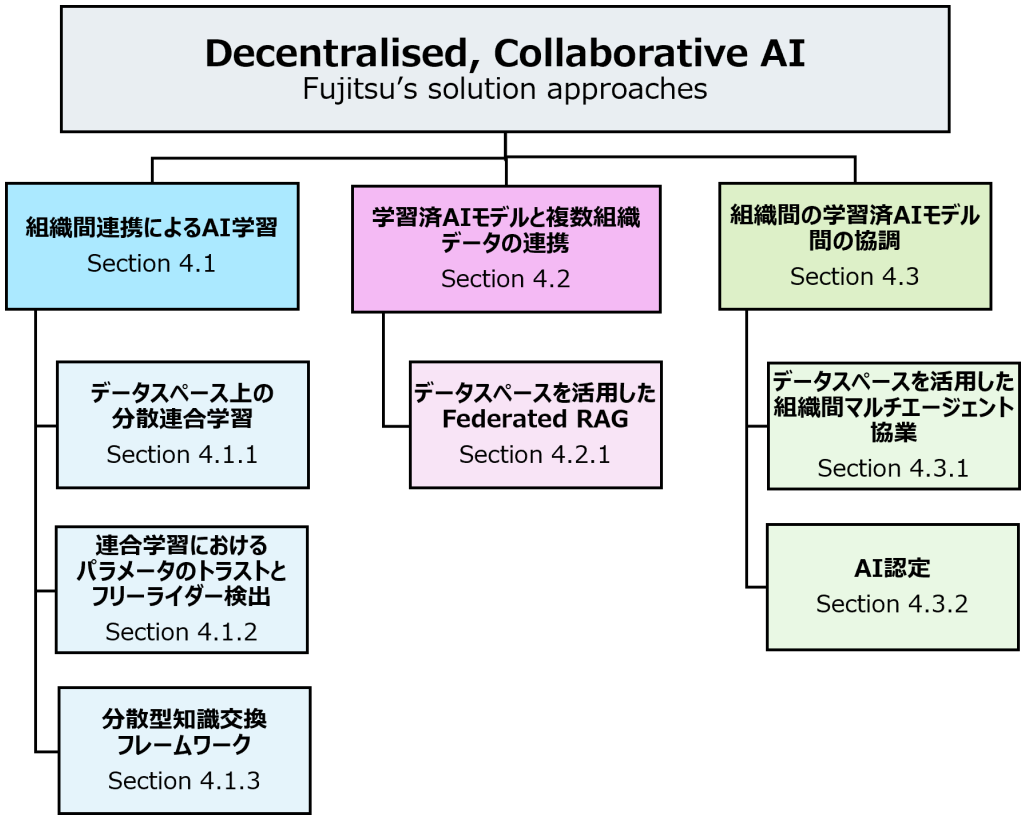

図 2 各々の類型に対する我々の取り組み一覧

4.1. 類型 1（組織間連携による AI 学習）の解決に向けて

4.1.1. データスペース上の分散連合学習

一般的な連合学習アーキテクチャでは、中央サーバーがすべてのローカル学習結果を収集し、そこから構成されたモデルを配布する役割を担います。しかし、この方式では、単一のインスタンスがモデルへのアクセス権限を付与する対象、目的、期間を決定します。さらに、参加者は自身の貢献を撤回できないため、共同モデルに提供した知識に対するコントロールを失うことになります。さらに、プライバシーは依然として複雑な課題です。過去の研究では、モデルの更新に基づいて元のデータが復元される可能性が示されており(Zhu et al., 2019, Nasr et al., 2019)、組織は貴重な知見を共有することに躊躇しています。組織が産業分野において連合学習を採用するためには、これらの課題をまず克服する必要があります。

富士通とフラウンホーファーISST は、プライバシーに配慮した主権型連合学習を実現するための新たな分散型アーキテクチャを開発することで、これらの課題に取り組むために協力しています。データスペース内に実装した予測保全のユースケースにおいてこのフレームワークの簡単な概要を示したものを図 3 に示します。このフレームワークの実現にはいくつかの課題が存在します。

連合学習には、モデルアンラーニングや準同型暗号のようなプライバシー強化手段など単体でも困難な課題がありますが、分散化することでさらに問題がさらに複雑になる可能性があります。我々の主権型連合学習フレームワークでは、この問題を解決しながら、以下の 2 つの核心的な要素を取り入れます：

1. エコシステムとガバナンス
   当社の連合学習フレームワークの基盤は、既存のデータスペース原則に基づいて構築され、分散型で安全かつ信頼性の高いエコシステムを提供します。これにより、分散型でモデル更新を交換・集約する仕組みを実現します。これには、公正な貢献インセンティブ、相互運用可能で自動的に強制可能な貢献ポリシー、およびネットワークトポロジーとモデル更新戦略に関する既存の研究を拡張し、主権型で分散型のモデル学習を促進する仕組みが含まれます。
2. 主権とプライバシー
   企業側が連合学習を完全に活用するためには、貢献者が貢献した知識に対する完全な可視性と制御を保持する必要があります。モデル学習中のデータプライバシーを向上させるため、モデル更新を共有する際の機密データ保護を可能にする新たなメカニズムを開発しています。また、完全なデータ主権を実現するため、モデル貢献者には、モデル貢献を撤回または取り消すための必要なポリシーと技術的メカニズムが提供されます。

4.1.2. 連合学習におけるパラメータのトラストとフリーライダー検出

連合学習では、データではなくパラメータを集約することで AI モデルを開発します。しかし、パラメータは単なる数値列であり、学習の元データとの関係が明確ではないことから、従来の真正性検証・品質管理の多くが適用できません。これに起因して、疑似パラメータを送信して報酬を得るフリーライダー問題などが発生します。

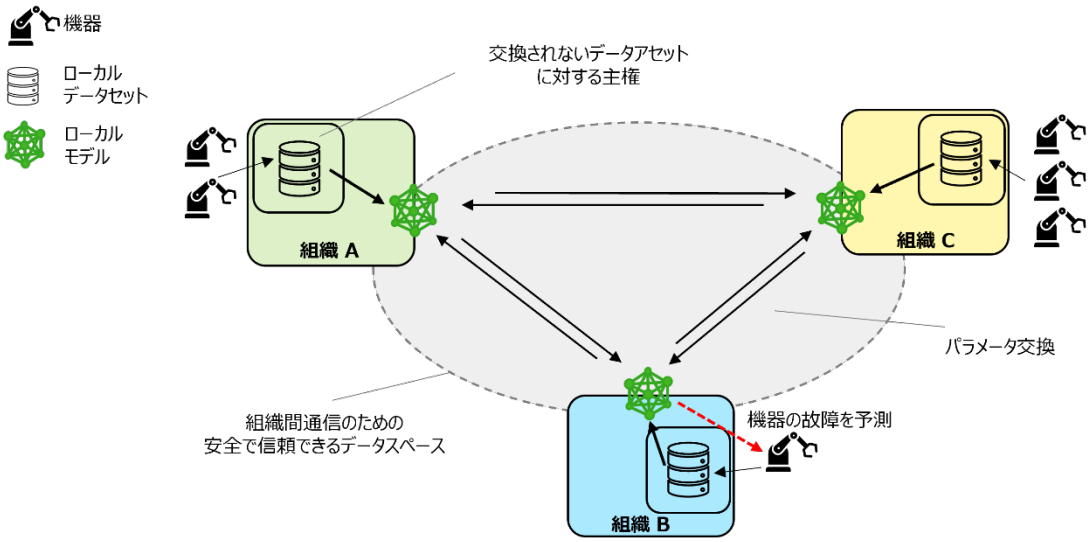

図 3 データスペースにおける分散連合学習 – 単純な予測保全のユースケース

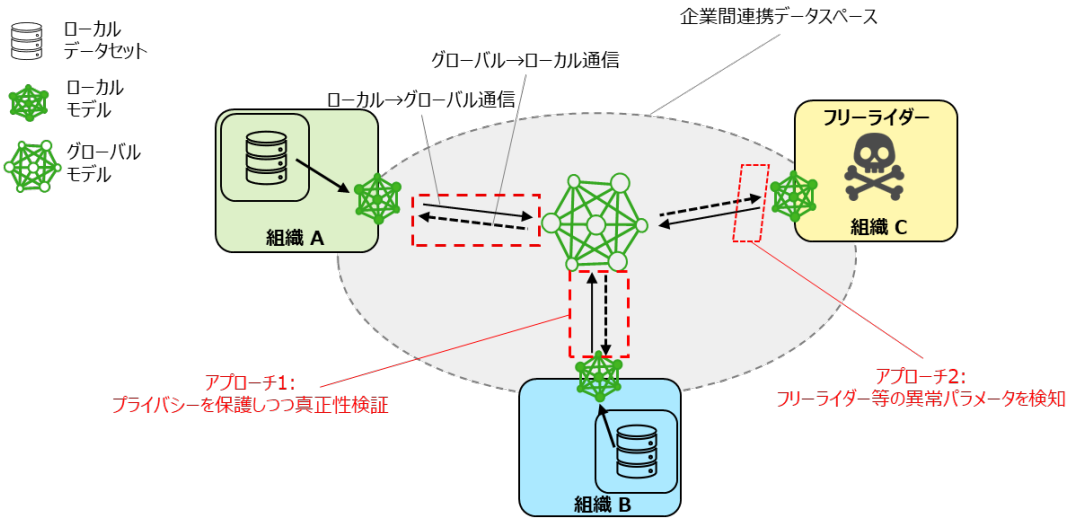

図 4 パラメータの真正性検証および品質管理

このような課題に対し我々は、パラメータの真正性検証・品質管理を実現するべく、暗号学的検証技術によるアプローチおよび異常検知技術によるアプローチという 2 つの側面から研究を進めています（図 4）。

暗号学的検証技術によるアプローチにおいては、AI へのゼロ知識証明応用研究（Chen et al., 2024; Abbaszadeh et al., 2024; Xing et al., 2023）を基礎として、パラメータの真正性を検証する技術を研究開発しています。特に、「クライアントが所持する生データにより学習された」ことを、生データを明かさずに検証することを可能にすることで、プライバシーとトラストを両立する技術を提案しました（福岡, 2025）。今後この技術をより実用的にしていき、クライアントとグローバルモデル（学習を集約するエンティティ）の間のパラメータ通信における双方向の証明・検証を実現することで、高水準な信頼関係を構築できる連合学習の具現化を目指します。

異常検知技術によるアプローチでは、フリーライダーをはじめとした学習を阻害しうるパラメータを検知・排除することで、パラメータの品質を管理する技術を研究開発しています。その一つとして、フリーライダーを抑止できる異常パラメータ検知技術を提案しました(中村 et al., 2025)。この検知技術は従来（Li et al., 2022; Cao et al., 2020）と比較して、グローバルモデル側のデータ収集コスト（Mazzocca et al., 2023）が小さいという特徴を持ちます。インセンティブが存在する連合学習においては、フリーライダー防止のニーズはより高まると考えられることから、我々はさらにこの研究を推進し、より高精度かつ低コストなフリーライダー検知技術に取り組んでいきます。

4.1.3. 分散型知識交換フレームワーク

従来の連合学習アプローチを通じた共同モデルの構築に、積極的に参加できない組織が存在する場合があります。これは、組織が自社のプライベートデータの価値を、共同モデルから得られると予想される利益よりも高く評価している場合や、他の組織の方が共同モデルからより多くの利益を得ると予想される場合などです。その他にも、組織がすでに独自のモデルを開発しており、そのモデルのアーキテクチャが提案された共同モデルと一致しないため、既存モデルを破棄するか再学習する必要があり、それに多大なコストがかかる可能性がある場合もあります。

こうした課題に対応し、できる限り多くの共同 LLM 学習のシナリオを考慮した、データスペースにおける AI モデル学習の包括的なアプローチを可能にするために、富士通は、モデルの更新を一切交換することなく、LLM 同士が完全に分散された方法で相互に学習する「分散型知識交換フレームワーク」を開発しています。中央集権型連合学習、分散型連合学習、分散型知識交換の 3 つのフレームワークの違いを説明する簡単な概要を図 5 に示します。

このフレームワークは、以下の 4 つの既存技術分野の概念を統合することで、参加全組織が共通の AI モデルを持つ必要がある従来の連合学習のボトルネックを回避します：

1. 分散／ピアツーピア型連合学習
2. パラメータ効率の高い LLM の連携
3. 協調型 AI におけるセキュリティ重視のガバナンス
4. データスペースおよびマルチエージェントの協調

さらに、この分散型知識交換フレームワークには、悪意のある参加者が結果を損なうことを防ぐための強固なセキュリティ保証が含まれています。

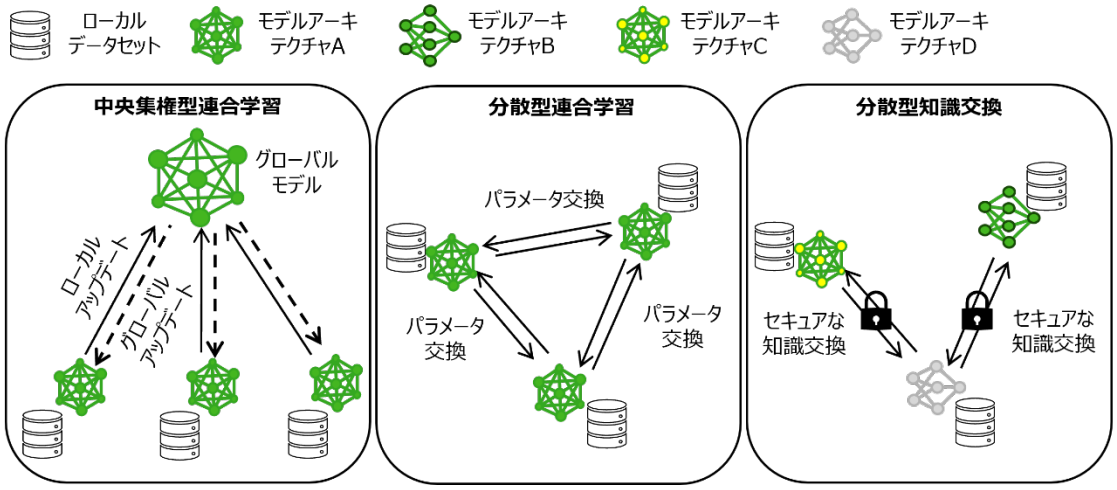

図 5 3 つの異なる分散型知識交換フレームワークにおける知識交換メカニズムの概要：中央集権型連合学習、分散型連合学習、分散型知識交換

4.2. 類型 2（学習済 AI モデルと複数組織データの連携）の解決に向けて

4.2.1. データスペースを活用した Federated RAG

RAG において、利用者の外部にある複数組織および企業のデータベースに問い合わせを行う Federated RAG （Seo, 2023)という手法があります。機密性や価値の高い企業データを使うことから、対象となるデータについては厳密な契約やアクセスポリシーを設定したうえで連携する必要があります。

我々は、ここでデータスペースの仕組みを拡張し、Federated RAG にデータ主権機能を追加する取り組みを進めています。この機能では、各組織にローカル LLM を実装することで、RAG の質問に対してどこからの問い合わせなのかを考慮して回答を生成したり、回答が利用者にどのように利用されるかを限定した回答を行ったりします。この仕組みにより、企業が持つデータに対する制御権の行使とデータアクセスを LLM 利用によって容易にすることの両立が可能になります。

これらの仕組み（図 6）により、データを持つ組織のデータ主権と利用者のデータ活用を両立させたデータスペースの実現を目指します。

4.3. 類型 3（組織間の学習済 AI モデル間の協調）の解決に向けて

4.3.1. データスペースを活用した組織間マルチエージェント協業

従来のマルチエージェントシステムでは、エージェントは単一のエンティティ（ユーザー）が指定した目標を達成するために相互連携します。タスクが他の組織がホストする外部エージェントに委任された場合でも、それらのエージェント

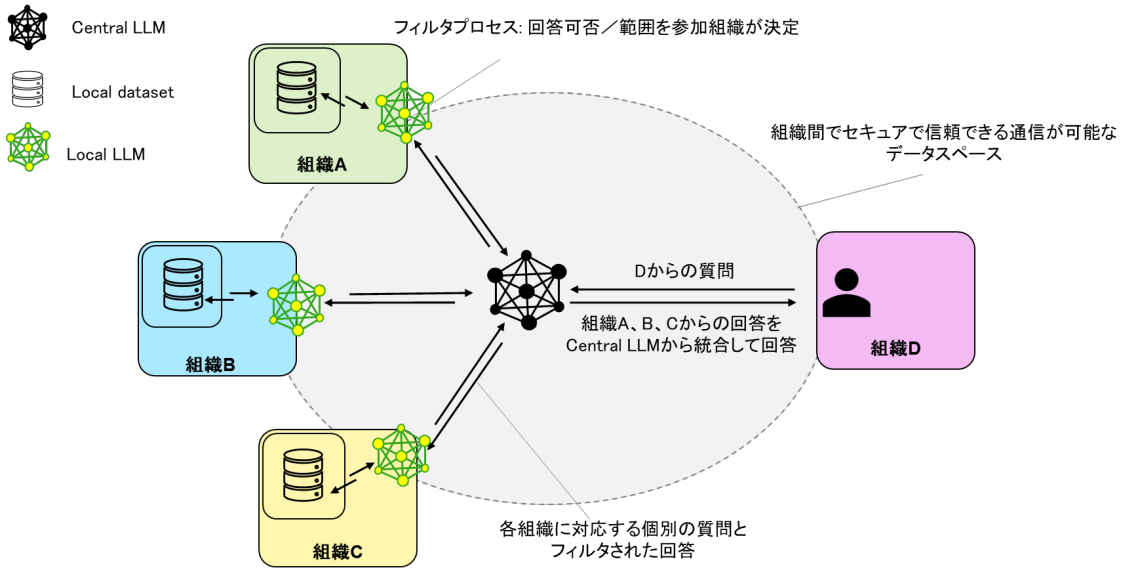

図 6 データ主権機能が追加された Federated RAG

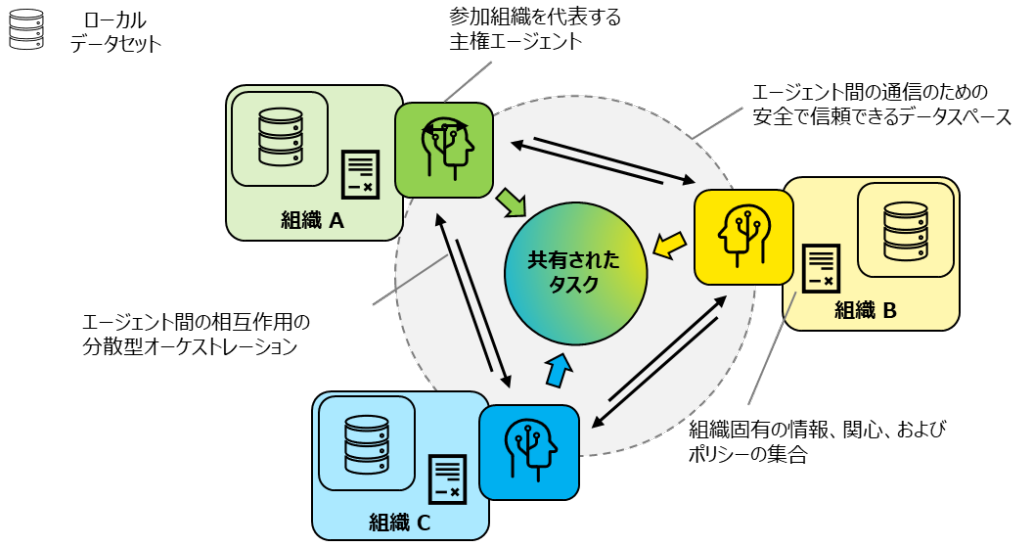

図 7 信頼性の高い組織間 AI エージェント協業を実現するための主権型マルチエージェントオーケストレーション

は通常、元のタスク提供者のリクエストに従って動作します。これに対し、我々の取り組む「組織間マルチエージェント協業」は、エージェントがより自律的に動き、利用元の各組織の個々の利益、好み、ポリシーを表明しながら、共通の目標を達成するためにエージェント間で協力できるようにします。そのために、この構想を実現するための様々な複雑な課題に対応していく、新たな主権型オーケストレーション手法が必要になります(図 7)。

第一に、エージェントは、提示されたタスクへの貢献の有無や方法を決定する自律性を持つ必要があります。エージェントが主権を持たない場合、通常はユーザーや自動化されたコントローラーが、タスクの解決にどのエージェントが貢献すべきか、およびその方法を決定します。しかし AI エージェントは個々の組織の利益を代表するものであるため、これらの組織の優先事項やポリシーがエージェント協業形成に向けて反映されることが重要で、リクエスト送信者や外部コントローラーからの命令で動くものではありません。これを実現するため、分散型エージェント発見と自律的協業形成のための新たなプロトコルを開発しています。

第二に、エージェント協業をホストするための透明性のある分散型フレームワークが必要です。信頼性があり、追跡可能で説明可能な意思決定プロセスを実現するため、エージェント間の相互作用は確立されたプロトコルに従い、中央主権原則を遵守し強制する通信チャネルを活用する必要があります。効率的な自律的意思決定には、例えば公正な投票メカニズム、議論履歴の記録、および個々の参加者の入力価値を判断するための貢献履歴追跡などの要素が不可欠です。私たちは、既存のデータスペース技術を拡張してこれらの機能を取り込んだ、組織間 AI エージェント協業のための信頼性が高く安全なプラットフォームを開発中です。

最後に、エージェント協業プロセスのオーケストレーションは、潜在的に対立する個々の利益や、信頼できない、悪意のある、または能力不足な参加者の脅威に対して頑強である必要があります。プライベートな利益は主権エージェントの相互作用において不可欠であり、適切に考慮される必要があります。一方、統計的には、能力不足の AI エージェントと遭遇することは十分想定されます。しかし、悪意のある、または有害な AI エージェントは即座に特定され、隔離される必要があります。この課題に対応するため、私たちはマルチエージェント協業向けの新たな粒度の細かい信頼フレームワークを構築しています。

4.3.2. AI 認定

3.3 節にて、安心して各組織が他組織の AI と連携する ためには、提供される AI モデルの信頼性を担保することが重要となり、そのための法規制やガイダンスの発行、認証制度などが開始されていることを述べました。また、この制度やプロセスに関する課題も説明しました。

特に、3 番目の課題として挙げた AI システムの認定の自動化を目指すにあたっては、認定機関だけでなくサプライチェーン上の各社も、AI システムに関する認定に必要な情報を自動的に蓄積するための証拠管理機能が重要となります。先行的な取り組みとして、AI-BOM(Bill of Materials)と呼ばれる、AI システムの構成要素であるモデル・データ・コード・インフラ等を網羅してリスト化して各企業で保持する仕組みも立ち上がりつつあります。ただし、現状ではまだ AI-BOM 単体を見ても AI 認定ができるわけではありません。

そこで、富士通は信頼性評価のための付加的情報の拡充に取り組みます。我々は、他組織のデータ共有におけるデータのトラストレベルを保証するための技術として Data LoA

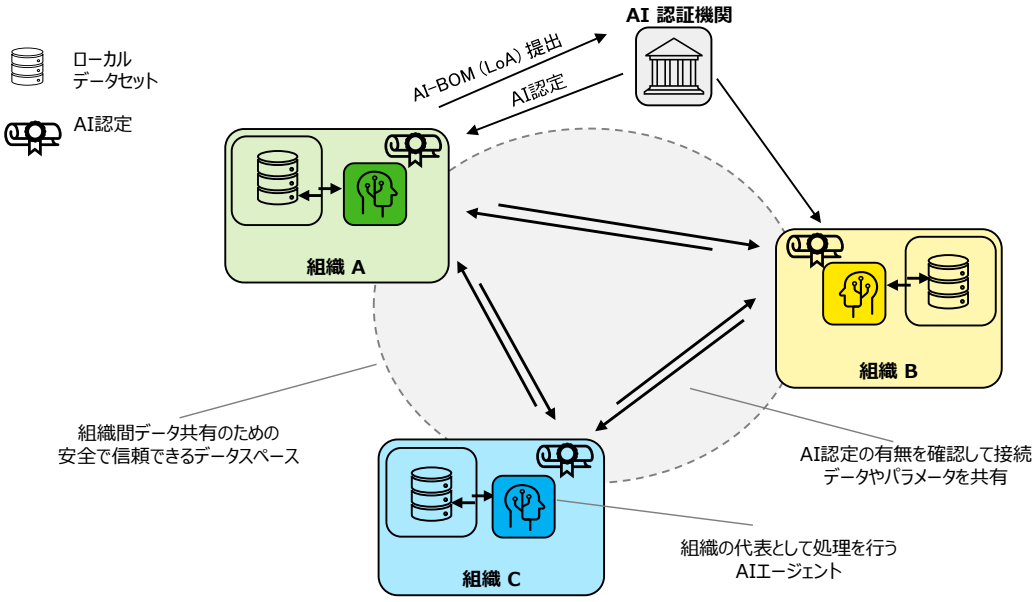

図 8 AI 認定を活用した組織間 AI 連携

(Levels of Assurance for Data Trustworthiness) (Fujitsu Limited, 2025; Zimmer et al., 2025) を開発しています。本技術を、AI-BOM に適用し、AI の信頼性を審査するときに、AI-BOM 自体の信頼性の評価を可能にすることで、AI 認定を促進します（図 8）。その他にも、AI の信頼性評価のために、さらにどのような情報をどのような形式で記録するのが良いかを検討し、安心した組織間の AI 連携の実現に貢献します。

5. 関連団体の動向と我々の技術提案

本ホワイトペーパーに記載したデータスペースを活用した分散 AI 技術は、関連する団体において、その基盤となる概念や課題意識に関する議論が進められています。我々は、こうした議論の流れを踏まえ、技術的な具体性を提示することで、分散 AI の社会実装に向けた議論の深化に貢献できると考えています。本章では、現在想定している関連団体の活動内容を整理し、我々がどのように貢献できるかを示します。

5.1. 産業競争力懇談会(COCN)

産業競争力懇談会(COCN)は、2025 年度の推進テーマとして「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, 2025)を立ち上げ、産業競争力の向上に貢献するため、AI 導入による価値創出やデータ基盤整備、エコシステム構築に向けた検討を進めています。

本推進テーマでは、生成 AI の進化に伴い注目される自律 AI エージェントを活用し、エンジニアリングチェーンとサプライチェーンの全体最適化を目指した取り組みが進められています。特に、異なる組織や現場から得られる多様な情報を横断的に活用すること、複数の AI エージェントが連携し、企業間の垣根を越えて工程の調整やリソース配分を自律的に行えるようにすることが重要なターゲットとされています。これらの取り組みにより、サプライチェーン全体の強靭性と効率性を高め、災害対応、人材不足、環境負荷への対応といった複雑な課題に対する柔軟な対応を可能にすることが期待されています。

本ホワイトペーパー 4 章で提示した技術は、企業間で連携しながら全体最適を目指していく点で、この COCN の推進テーマと高い親和性を有しています。類型 1 の技術を活用した、サプライチェーン上の企業間におけるデータ連携による AI 学習による在庫管理や設備保守、またこの AI 学習に参加するためのインセンティブモデルやフリーライダー防止策の提示、類型 2 の技術による自然言語を活用した柔軟な組織間の情報連携、類型 3 の技術による各組織が持つ様々な AI エージェントの連携、などです。我々は、これらの技術を積極的に本推進テーマおよび関連活動に提案していきます。

5.2. 日欧データスペースイニシアティブへの貢献

本ホワイトペーパーでは、データスペース上で実現する分散協調 AI の提案を行っています。データスペース関連の取り組みは日欧双方で活発です。日本においてはウラノスイニシアティブや DATA-EX、欧州においては、Catena-X や Gaia-X、IDSA などが挙げられます。これらの取り組みにおいては、データスペースに関するさまざまな標準を作っていくとともに、オープンソースソフトウェア(OSS)として実装を行っています。

本ホワイトペーパーで提案する技術群は、データスペース上での活用を想定しています。そのため、関連するデータスペースの標準や、その実装としての OSS に貢献することで、データスペースにかかわるステイクホルダーが容易に技術にアクセスし活用できるようにしていく予定です。

6. おわりに

本ホワイトペーパーでは、データスペースにおけるプライベート AI の実現に向けた概念と技術の方向性について議論してきました。企業が競争優位性を確立し、高度な意思決定を行うためには、組織固有のデータを活用した AI が不可欠です。しかし、単一企業が持つデータには限界があり、企業間でのデータ連携が重要となります。それに対し、データスペースは、データ主権を軸とした安全なデータ共有基盤として、企業間連携を促進し、より高度な AI 開発・活用を可能にします。本ホワイトペーパーでは、データスペースにおけるプライベート AI 実現に向けた課題と、その解決に向けた富士通の取り組みを紹介しました。今後は、本ホワイトペーパーで提示した技術を、日欧をはじめとした様々な場に展開していく予定です。

富士通は、データスペースにおけるプライベート AI の実現に向けて、今後も技術開発と実証実験を推進し、企業がデータに基づいた意思決定を行い、新たな価値を創造できるよう支援していきます。

参考文献

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.

本件に関するお問い合わせ

富士通株式会社

富士通研究所　データ＆セキュリティ研究所

Email: fj-aifordataspaces@dl.jp.fujitsu.com