NTT DATA

The Convergence of Data Spaces and AI

A Strategic Opportunity for Europe

Executive Summary

Artificial Intelligence and data spaces are converging at a strategically important moment. Data spaces were designed to enable trusted, sovereign and interoperable data sharing, and AI is reshaping how data is discovered, used and monetised. They can create a new opportunity together, where data spaces can provide the high-quality data that AI increasingly requires, and AI can make data spaces more usable, interoperable, valuable and sustainable.

This convergence is particularly relevant in the European context. Europe is building Common European Data Spaces and a middleware for European data spaces to support data sharing under European rules and values. Data spaces are already a part of the differentiating foundation for European AI, especially in important domains such as health, mobility, manufacturing, energy and public administration, where trust, traceability and legal certainty matter as much as scale.

The convergence of AI and data spaces also introduces new challenges and risks. Governed access to data does not automatically make data suitable for every AI purpose. Generative AI can weaken trust if synthetic or AI-assisted content is not properly identified and validated. Usage restrictions such as “no training” remain difficult to enforce. AI can increase the risk of data escaping the original sovereignty boundaries, and agent-based interaction raises new questions around accountability.

This paper explores the transition and its aspects and challenges.

Current Context: Two Paradigms Converging

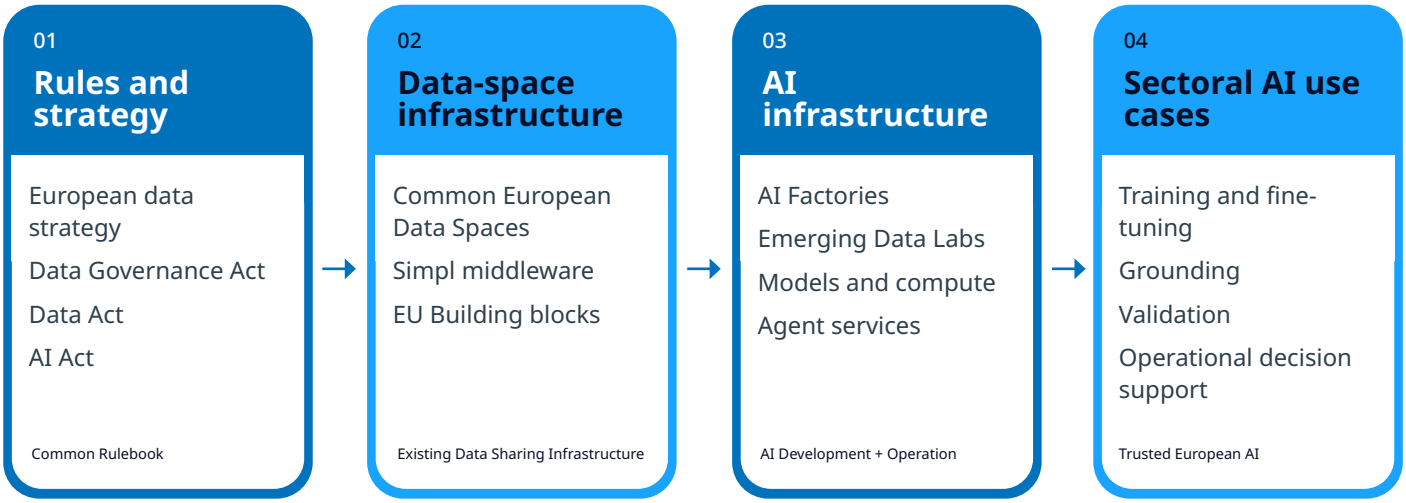
Data spaces and AI are now converging in a way that is strategically important for Europe. Data spaces enable secure, trustworthy and sovereign data sharing across organisational boundaries, in line with the European strategy for data and sovereignty. For example, Common European Data Spaces are intended to make high-quality data available for access and reuse in a trustworthy and secure environment, preserving European values on how data is governed and used. This is closely linked to the Union's broader objective of strengthening competitiveness, innovation and data sovereignty[1].

This policy and legal architecture become even more relevant with the rise of AI. The value of advanced AI increasingly depends on access to high-quality and legally reusable data, and in sensitive or high-value sectors, the need is even stronger. The European Health Data Space is a clear example that creates a cross-border framework for access to and use of electronic health data, supporting healthcare delivery, research, innovation and public-interest.

The convergence is also becoming visible in the evolving European policy and infrastructure landscape. Recent EU initiatives such as AI Factories, and the emerging concept of Data Labs, show that Europe increasingly sees trusted access to high-quality data as a critical enabler and prerequisite for AI development. This convergence is a strategic opportunity for Europe: it is building a rich ecosystem of data and data-sharing practices that can feed AI training, and AI developers in other geographies are unlikely to have access to a data landscape of comparable depth.

At the same time AI is changing the way data is discovered, interpreted and consumed. Large language models, retrieval-based systems and AI agents are introducing more conversational, contextual and automated ways of interacting with data. In this context, grounding refers to connecting an AI model to relevant external data, for example data retrieved from a data space, so that its outputs are informed by trusted, context-specific information rather than relying only on what the model learned during training.

Traditional data space interaction models are still often organised around catalogues, metadata, access conditions and negotiated exchange. Those models do require human involvement, especially in governance, supervision and accountability. However, the underlying data-contracting, data transfer and processing can already take place machine-to-machine. What AI can change is that discovery, selection, contract negotiation support and operational decision-making can increasingly be automated.



[1] European strategy for data; Common European data spaces; Data Governance Act, Regulation (EU) 2022/868; Data Act, Regulation (EU) 2023/2854

Mutual Reinforcement Between AI and Data Spaces

The relationship between AI and data spaces is bidirectional and goes across different stages of their lifecycle. Data spaces can improve the quality, legality and trustworthiness of the data used by AI systems, while AI can improve the way data spaces are operated, understood and consumed. This understanding matters, because it moves the discussion beyond a generic claim of “synergy” and shows more precisely where value is created.

For many AI use cases, especially in regulated or high-value sectors, the challenge sometimes is to access data that is relevant, well-described, provenance-rich and reusable under clear conditions. This is closely aligned with the logic of the EU AI Act, which places particular emphasis on data quality and data governance for high-risk AI systems, and with the broader European data strategy, which aims to make more data available for reuse through trusted and interoperable data-sharing mechanisms. In this sense, data spaces offer a disciplined alternative to indiscriminate data aggregation.

The value of data spaces for AI lies in the combination of quality, legal clarity and sector-specific depth. In practice, this means that data spaces can support forms of AI development that are more traceable, more policy aware, and ultimately more robust and resilient over time, thanks to the availability of high-quality data that can be governed, maintained and reused over time. This emphasis on resilience is also consistent with the broader European regulatory landscape, including frameworks such as NIS2 and DORA, which reinforce the importance of cybersecurity, risk management and digital operational resilience.

AI strengthens data spaces in a different way; it can reduce the friction that limits interoperability and adoption. One of the recurring difficulties in data spaces is that alignment is rarely only technical. Participants may use different schemas, vocabularies, classifications and documentation practices even when they broadly agree on standards. AI can assist here by supporting semantic mapping, metadata enrichment, translation between partially aligned models, and detection of inconsistencies or gaps. In that sense, AI can become an interoperability accelerator.

AI can also change the interaction model of data spaces. Traditional data spaces are often experienced through catalogues, forms, and predefined dashboards. AI can make it possible for users to express intent in natural language and receive not only search results, but also synthesised answers, tabular views, visualisations and reports generated on demand from data sources. This lowers the threshold for non-technical users, so they no longer need to understand the structure of the underlying catalogue or data model in detail and reduces the need to pre-build and maintain every chart, report or workflow in advance. AI can turn data spaces from static discovery environments into more responsive decision-support environments, which changes how value is extracted from the ecosystems.

A further step in this evolution is the emergence of AI agents as operational intermediaries. In a mature setting, an agent can act on behalf of an organisation within clearly delegated authority. It may search for relevant assets, prepare or initiate contract negotiations, request access under defined policies, trigger processing workflows, or route results into downstream systems. AI agents can extend the usability of the Data Space Protocol (DSP) by increasing the degree of automation.

Once data space interaction becomes more agent-driven, consumption and billing models will also change accordingly. In an AI-enabled model, data and services are more likely to be consumed per API calls, per model-grounding requests, processing jobs, or contract period. This can improve the sustainability model of Data Spaces, which is placed by the DSSC Blueprint as a core concern of data spaces.

DATA SPACES → AI

Data spaces strengthen AI

- Curated, sector-specific data
- Provenance and traceability
- Clear access and reuse conditions
- Domain depth and continuity

DATA + TRUST

USABILITY + VALUE

AI → DATA SPACES

AI strengthens data spaces

- Agent automation and new services
- Metadata enrichment
- Natural-language interaction
- Semantic mapping

Risks and Challenges at the AI-Data Space Interface

The convergence of AI and data spaces creates risks in both directions. AI can weaken some of the trust assumptions on which data spaces depend, while data spaces can create a misleading impression that governed access automatically makes data suitable for any AI purpose. The central challenge is preserving enforceability and accountability when data is generated, combined and consumed by machines.

1. Provenance, authenticity and the arrival of synthetic content

One of the most immediate risks is the growing presence of AI-generated or AI-manipulated content within data ecosystems. In many domains, generative systems can already produce text, images, structured records, code and even metadata, which can be useful for simulation, augmentation, anonymisation support, testing or documentation. Challenges arise when the artificial origin of the data and the validation status remain unclear.

In specialised sectors, this becomes particularly serious. Where there is limited ground truth, plausible-looking outputs may be difficult to verify, and the apparent sophistication of generated content can mask factual weakness. In such cases, poor provenance becomes a trust problem for the entire data space. A participant may no longer know whether a dataset is directly observed, statistically derived, AI-assisted, or fully synthetic, nor under what controls it was validated.

Article 50 of the AI Act introduces transparency obligations for certain AI-generated or AI-manipulated outputs, and Commission guidance explicitly refers to machine-readable marking, detection mechanisms, metadata, watermarks and cryptographic techniques as ways to support transparency. Therefore, data spaces should be able to represent whether content is AI-generated, or AI-assisted, and under what process it was reviewed or validated.

2. Governed data is not automatically AI-ready

Data spaces improve governance, documentation and access control, and it is most of the times expected that the data there should be of high quality. Therefore, it may create the impression that data obtained through data spaces is automatically suitable for AI development.

A dataset can be controllably shared and well described, yet still be inadequate for training, validation or testing of an AI system. It may be incomplete, too narrowly scoped, biased or unsuitable for the intended context of use. This point is important within the AI Act, since it does not equate data availability with data suitability. For high-risk AI systems, the Act requires training and validation datasets to be relevant, and as far as possible free of errors. In other words, participation in a trusted data space may improve the starting point for AI, but it does not remove the need for rigorous assessment of the quality and fitness for purpose.

3. Enforcing “no training” and other usage restrictions in AI pipelines

One of the biggest challenges at the intersection of AI and data spaces is the enforceability of usage rights. Similar to real data spaces, in AI systems there is no easy way to validate whether policies and usage rights are being fully respected. Policy languages can express such restrictions, and contracts can reinforce them, but technical enforcement remains difficult. Models can behave differently, datasets can get copied, or fully or partially regenerated without proper transfer of their usage rights, and sensitive data that is marked for “no training” may end up in ML pipelines for various reasons.

Recent EU measures improve transparency, but they do not solve this problem. This is still a big operational problem, particularly where confidential business data or personal data are involved.

4. Inference, recombination and loss of sovereignty over derived knowledge

Data sovereignty risks exist beyond raw data access. Specifically, AI systems can combine lawfully obtained datasets and infer new knowledge that no single participant intended to disclose on its own. This is one of the most underappreciated challenges in AI-enabled data spaces.

Consider a situation where a participant may share operational, environmental, maintenance or commercial data under carefully negotiated conditions. Yet once such datasets are combined and analysed by advanced models, they may reveal supplier dependencies, production patterns, business vulnerabilities, and behavioural profiles or other sensitive derivatives. From a legal and commercial perspective, the difficult question is not only “who accessed the data,” but also “who now controls the insight generated from the data.”

In AI-enabled ecosystems, sovereignty matters also over derived artefacts and analytical outputs generated from combined sources. If governance remains focused only on the raw dataset, it may fail to protect the actual value or sensitivity that AI extracts from it.

5. Agent-mediated access, accountability and cost exposure

A further challenge emerges when AI agents begin to act on behalf of organisations inside data spaces. Agent-based interaction promises efficiency, but it also introduces new questions of delegation, liability and control.

An agent may be authorised to discover data, initiate negotiations, trigger processing, or obtain access under predefined conditions. But if that agent acts at machine speed across several services, the traditional governance assumptions of manual oversight become weaker. Who authorised the action? Under what spending limit? Against which policy version? With what right of onward use? An agent can over-consume a priced service or triggers excessive downstream processing.

This is why identity and policy may no longer be sufficient, and AI-enabled data spaces will also need strong delegation models, auditable execution logs, consumption metering, spending controls and settlement mechanisms.

Strategic Direction: Toward AI-Native, Trust-Preserving Data Spaces

The convergence of AI and data spaces should not be approached as a simple layering exercise, where AI is added on top of existing data-sharing infrastructure. The stronger strategic direction is to treat this convergence as a redesign challenge on both sides.

1. From "AI in data spaces" to "data spaces for AI"

A useful starting point is to avoid defining the ambition too narrowly, for example to use AI to search catalogues, or to automate some functions. The real opportunity is that data spaces can change the way AI is developed. In the current global AI landscape, competitive advantage is often associated with scale of compute and scale of training data. Europe's stronger position may lie in combining high-value sectoral data, trustworthy access conditions, provenance, and legal certainty.

This means that the differential value of European AI may come not only from training larger models, but from training and fine-tuning models on, and grounding their outputs in, data that is better curated, governed and more closely aligned with real sectoral needs. Europe can pursue AI that is more trustworthy, domain-credible and more compatible with its regulatory model.

2. Move from catalogue-centric access to AI-native interaction

If data spaces are to remain relevant in this new environment, they need to evolve beyond the assumption that value is created mainly through human users browsing catalogues and manually negotiating access. That interaction model will remain available for governance, accountability and oversight, but it will no longer be sufficient as the primary mode of operation. A more suitable target is an AI-native interaction model in which data spaces expose their capabilities for both humans and AI Agents.

This is also where interaction design matters more than it might appear. Natural-language access and context-aware guidance can make data spaces substantially more useful, especially to non-technical participants. This can widen participation, reduce the need to pre-build large numbers of fixed dashboards, and lower maintenance effort over time. This can also improve the sustainability of the data spaces, by broadening the potential end-users.

3. Make policy, purpose and provenance technically actionable

The risks identified earlier point to a new requirement: If AI-enabled data spaces are to be credible, they need technical mechanisms that carry legal and governance intent into runtime operations.

Data spaces should be able to represent not only where data came from, but how it was transformed, under which conditions it may be reused, and whether it contains human-generated, AI-assisted or AI-generated elements. That direction is consistent with the transparency logic of the AI Act and with the broader need to preserve trust in mixed data environments. It should also extend to the AI lifecycle itself: organisations need to distinguish datasets suitable for training from datasets suitable only for controlled processing. The fact that data is shared lawfully in a data space does not, by itself, determine that it is appropriate for every AI purpose.

From there, the more difficult step is enforceability. Usage conditions such as "no model training," "no onward sharing," or "processing only within a trusted execution environment" must increasingly be supported by technical controls rather than by contract language alone. Secure processing environments, policy-aware orchestration, controlled export, and hardware- or platform-assisted isolation can all contribute here, depending on the sensitivity of the data and the intended use case.

4. Design for delegated agents and machine-mediated exchange

A further strategic shift is needed around actors and identities. Once AI agents begin acting on behalf of organisations, the data space can no longer assume that every meaningful action is directly initiated by a human operator. Agents may search, compare, request, negotiate, analyse and trigger downstream workflows at speeds and frequencies that traditional governance models were not built for.

This does not mean that data spaces should accept agent autonomy uncritically, instead, they should design for it explicitly. Agent-mediated participation will require stronger delegation models, clearer scoping of authority, auditable decision trails, and technical identity mechanisms capable of linking agent actions back to accountable organisations. In governance terms, the relevant question becomes not only “who is the participant?” but also “which agent was acting, under what mandate, under which policy version, and with what spending or usage limits?” The answer must be observable after the event.

5. Treat metering, billing and settlement as core capabilities

Finally, AI-native data spaces will need a more mature economic layer than current ecosystems require. Once data processing and agent-driven interactions become more dynamic, coarse models based only on membership or static access rights become less adequate. Consumption becomes finer grained and more variable: an organisation may pay for data access, a secure computation session, a semantic mapping service, or a chain of automated operations initiated by an agent.

European initiatives such as Simpl are moving to that direction, by defining various monetisation capabilities for the open-source middleware for data spaces.

Sustainable data spaces need auditable usage events and links between policy terms and chargeable events, and a settlement model that participants can understand and trust. This is fully aligned with the DSSC Blueprint, which treats business, governance and legal building blocks as essential for financially sustainable and value-driven data spaces.

Conclusion

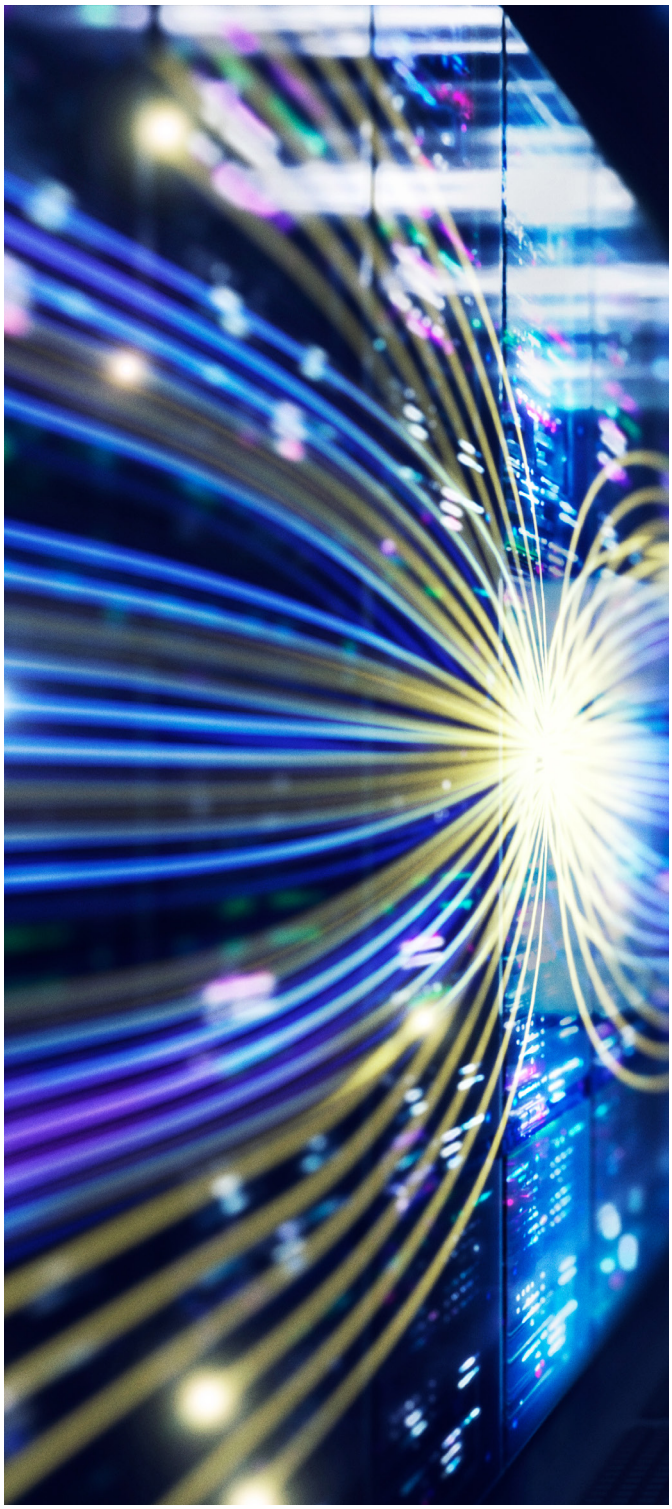
AI and data spaces are no longer parallel developments. They are becoming part of the same emerging architecture of digital value creation. Data spaces offer AI something increasingly scarce and strategically important: access to high-quality data. AI, in turn, can help data spaces move beyond static discovery and manual interaction toward more dynamic, interoperable and usable ecosystems.

Trust cannot be assumed simply because data is shared within a governed environment. Data that is accessible is not necessarily suitable for training. Policies that are defined are not necessarily enforceable. Sovereignty over raw data does not automatically extend to derived insights. And agentic automation that improves usability can also create new forms of accountability and cost exposure if agent behaviour, consumption and delegation are not properly controlled.

If data spaces are connected effectively with AI development, grounding and operational use, they can help shape a distinct European approach to AI, with focus on traceability, lawful reuse and sectoral credibility. That may become one of Europe's strongest differentiators in an AI landscape often dominated by scale-first models.

The direction is therefore clear: Data spaces need to evolve from catalogue-centric exchange environments into AI-native, policy-aware and economically operable ecosystems. This requires stronger provenance, clearer distinctions between different AI uses of data, technical enforcement of usage restrictions, governance for delegated agents, and auditable metering and settlement models. If these elements are treated as foundational rather than optional, data spaces can become more than enablers of data sharing; they can become part of the trusted infrastructure for European AI.

Looking ahead, the next step is to scale AI-native data spaces and strengthen their integration with AI ecosystems. Practical mechanisms for provenance, usage control and agent governance will be essential for data spaces to become part of Europe's trusted AI infrastructure.




NTT DATA