AI Factories and the data challenge: access, acquisition and usage of data. Connection to data spaces

Discussion paper

Draft version 21-10-2024

Table of Content

Context and purpose

The present document is a discussion paper written by BDVA in collaboration with its members (including some of the future AI Factories hosting organisations) and some Common European Data Spaces. **It focuses on highlighting the value and opportunities that arise from the connection of AI Factories and EU Data Spaces, identifying the requirements for this connection and providing some guidance on how to address existing challenges.** It is a living document that intends to build coherence, common knowledge and to propose recommendations and a roadmap of actions to accelerate this connection.

BDVA is a private member of the EuroHPC JU, representing the role of user industries and research on data and AI. BDVA is very active in the EU Data Space landscape, partner in the Data Spaces Support Centre and collaborates closely with other key initiatives on the data spaces domain (notably the Data Spaces Business Alliance, established by Gaia-X, IDSA, FIWARE and BDVA). BDVA members are key contributors to the Data Space programmes at European and National level. Additionally, BDVA and its members are very active on Industrial AI and Data-driven AI (for public, personal and industrial usage); in this context BDVA contributes to the objectives of the ADR partnership and collaborates with Adra and other stakeholders in Europe. BDVA and its members have also gained over the years a lot of experience engaging AI startups and SMEs, mainly through data incubator programs (targeting secure data sharing and AI innovators) and i-Spaces[1]. BDVA i-Spaces contribute to EDIHs, AI TEFs, AI NoEs and HPC CCs, often functioning as a comprehensive 'one-stop shop' for the data and AI ecosystem at national and regional levels. Moreover, many of the future hosting organisations of AI Factories are labeled i-Spaces or members of BDVA, mainly connected to BDVA for Big Data and AI research and innovation activities under EuroHPC.

The launch of AI Factories, their need to access a massive amount of data and the value to engage with Common European Data Spaces, and the identification of requirements and support to user engagement, are at the core of BDVA objectives and activities. In this context, BDVA is well positioned to offer support to the EuroHPC JU, EC, AI Factory hosting organisations and EU Data Spaces in finding an effective path towards a successful implementation of AI Factories based on user-focused offerings for industry, SME and start-ups.

Collaborations with all stakeholders have been established, but increasing collaborations and more hands-on knowledge will be needed so we expect this document to be enriched with additional contributions in the upcomign months

Note: **This document is not intended for public wide distribution in its current version**. It is planned to open this document to a much wider audience once the maturing levels of the content increase.

[1] https://bdva.eu/ispaces/

1. Background

AI Factories (AIFs), a cornerstone of the AI Innovation Package launched by the European Commission in January 2024, are expected to play a pivotal role in strengthening Europe's leadership in trustworthy AI. Designed as open ecosystems centered around European supercomputers dedicated to AI, these AI Factories integrate high-performance computing, access to data, and expertise to support AI-driven SMEs, startups, research and innovation ecosystems, and other productive sectors in the development and training of large AI models[2]. The significance of AI Factories has been underscored in Dragi's report, and European Commission President Ursula Von der Leyen has charged Commissioner-designate Henna Virkkunen in her mission letter with the task of ensuring tailored supercomputing capacity for AI startups and industries via AI Factories within the first 100 days of her mandate. The first two calls for the selection of entities to establish AI Factories have been published by the EuroHPC JU[3] last 10th September 2024.

Data is fundamental for the development and training of large AI models. To accomplish their objective, AI Factories should be able to access, acquire and reuse large amounts of different types of high-quality data, spread out in the cloud and fragmented through many repositories and data ecosystems, and coming from different sources. Additionally, this data might be subject to different conditions and terms of access, use and re-use, and specific regulations may apply to different datasets and data ecosystems, particularly important for both privately owned (proprietary data) and personal data. Finally, AI Factories must be able to guarantee the protection and security of sensitive data, and to provide secure and trusted environments so results from their activities are compliant with current regulations, especially the AI Act.

The **specific data requirements to the AI factories** are explained in the text of the recently published calls (section "Description of an AI Factory" and "Annex: AI Factories concept paper"), and can be listed as follows:

The availability and accessibility to large data repositories with high quality curated data is fundamental for the AI community to flourish. AI Factories must guarantee high-speed connectivity and unrestrained access to European Data Spaces and relevant data repositories.

- ***Data facility:*** *Co-located or very high-speed connection to (at least) one associated data facility linked to the supercomputer. Data centres to host large volumes of data necessary for AI Factories and associated data facilities must be operational within 12 months of being selected to host an AI Factory.*
- ***Access to Common European Data Spaces[4]:***

  Hosting entities should clearly identify interaction with and access to which Common European Data Spaces they wish to interact and have access to, provided that these correspond to their targeted / selected applications / domains that are aligned with the strategic vision and strategic specialisation areas of the hosting country / hosting Consortium. Hosting Entities should also describe the principles of an eventual access to and use of agreement with such Common European Data Spaces.

[2] https://digital-strategy.ec.europa.eu/en/library/communication-boosting-startups-and-innovation-trustworthy-artificial-intelligence

[3] https://eurohpc-ju.europa.eu/supercomputers/selection-hosting-entities_en

[4] Common European Data Spaces | Shaping Europe's digital future (europa.eu)

Complementary and relevant data repositories (e.g., Hugging Face) should also be considered, as well as readiness to integrate into the future EuroHPC Federation Platform, which will be federating EuroHPC JU supercomputers and European HPC resources.

- ***Security.*** *AI Factories should guarantee the confidentiality and integrity of sensitive data and ensure the integrity of computational processes. Users of computing capacity could for example be authenticated using the EU eID Wallet, once available.*
- ***Secure and Trusted environments.*** *Where justified, AI factories should establish secure and trusted (research) environments for both industry and scientific research ensuring the confidentiality and integrity of data*

In this regard, **EU Data Spaces** emerge as ideal instruments to establish a mutual beneficial relationship with AI Factories. EU Data Spaces are interoperable frameworks of common standards and practices, to share or jointly process data, for the development of new products and services[5]. These interoperable frameworks, based also on common governance principles and enabling services, are also aimed to enable trusted data transactions between participants[6]. EU Data Spaces aim at overcoming legal and technical barriers to data sharing across organisations, by combining the necessary tools and infrastructures and addressing issues of trust[7]. EU Data Spaces can support AI Factories on addressing the diversity of sources, data and conditions, provide governance mechanisms to ensure data protection and sovereignty, and support compliance to data regulations, as well as provide data interoperability, security, governance and quality. On the other hand, AI Factories might complement EU Data Spaces with HPC resources, access to additional data, specific AI and HPC services, and access to large AI models already trained or pre-trained.

This document explores this win-win relationship between AI Factories and EU Data Spaces, highlighting their common value proposition and potential opportunities, but also providing some guidance on how to start this collaboration, identifying the most upraising relevant challenges and recommendations to overcome them.

[5] https://digital-strategy.ec.europa.eu/en/policies/data-act
[6] https://dssc.eu/space/bv15e/766061638/1+Key+Concept+Definitions
[7] https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_en

2. Mutual value generation and opportunities

As highlighted in the previous section, AI Factories and EU Data Spaces can mutually benefit each other and jointly bring a differentiating factor to Europe, unlocking the power of European data and supercomputers to propel EU AI innovation, and eventually benefiting all parties involved in AI Factories and data spaces, particularly small actors such as SMEs, start-ups, innovators and researchers.

The EU has a significant opportunity to make an impact by unlocking data that is currently not publicly accessible—especially industrial data. This aligns with one of the primary goals of data spaces: t**o provide access to private industrial data under well-defined conditions and usage policies, offering a major advantage to AI innovators and enhancing the mission of AI Factories**.

Moreover, **EU Data Spaces bring together large industrial communities**. These companies have already joined forces and established collaborations around data spaces, to provide and share data in order to generate value for their businesses. The connection of AI Factories with data spaces will also connect them to these industrial communities (e.g. to develop specific industrial foundation models, etc.), attracting more industrial users to them, and solving the traditional unawareness of SMEs/startups and other organisations about the availability of HPC services in HPC CC. Therefore, AI Factories can benefit from the ecosystems already created around data spaces, to raise awareness of the HPC CC capacities and their offer. Additionally, this can also accelerate the value creation within data spaces.

More specifically, users of AI Factories may bring their own data to train their models, which can be stored and made accessible through local data facilities. However, the true advantage of the AI Factory concept lies in its ability to enrich these models with sector-specific industrial and high-quality data that was not initially available, made possible through its connection to data spaces

Data spaces can also provide the flexibility needed to have the data prepared to the level required by AI practitioners. While the power of data spaces relies on the availability of massive amounts of data, they can also provide services to prepare, process, curate and filter data to a certain extent. Data spaces can be required to complement the data with specific data models and metadata descriptions specially aimed for ML / AI (e.g. Croissant[8] vocabulary or similar), document data (Datasheets for Datasets) or complete data quality description to be compliant with AI Act (all these aspects explored in BDVA discussion paper “Elevating Data Quality: A Paradigm Shift for Data Spaces and AI Needs”).

Data spaces also offer **controlled and well-governed environments to ensure proper governance and compliance mechanisms for the data they provide**. While additional technical measures may be required (as discussed in Section 4), these can be extended to cover the entire data lifecycle within the AI Factory, enabling end-to-end data governance and compliance. This allows the AI Factory to leverage the existing framework provided by the data space, rather than starting from scratch.

[8] https://github.com/mlcommons/croissant

Based on the above, it is clear that AI Factories provide a well-defined purpose for EU Data Spaces, positioning themselves as powerful and relevant use cases, and potentially becoming some of their first major customers.

But AI Factories can also complement the offerings of data spaces by providing access to HPC infrastructures when required by the business model or use cases of the data space, along with the necessary HPC services and support. As such, AI Factories can be viewed as an extension of the hardware capacity within data spaces. As part of its AI dimension, the AI Factory can also act as a service provider in the data space, with services related to AI, access to additional datasets and to the library of AI models already trained or pretrained in the AI Factory (subject to the specific licenses and usage policies).

Ultimately, all the above benefits the end users of AI Factories and data spaces (see Figure 1).

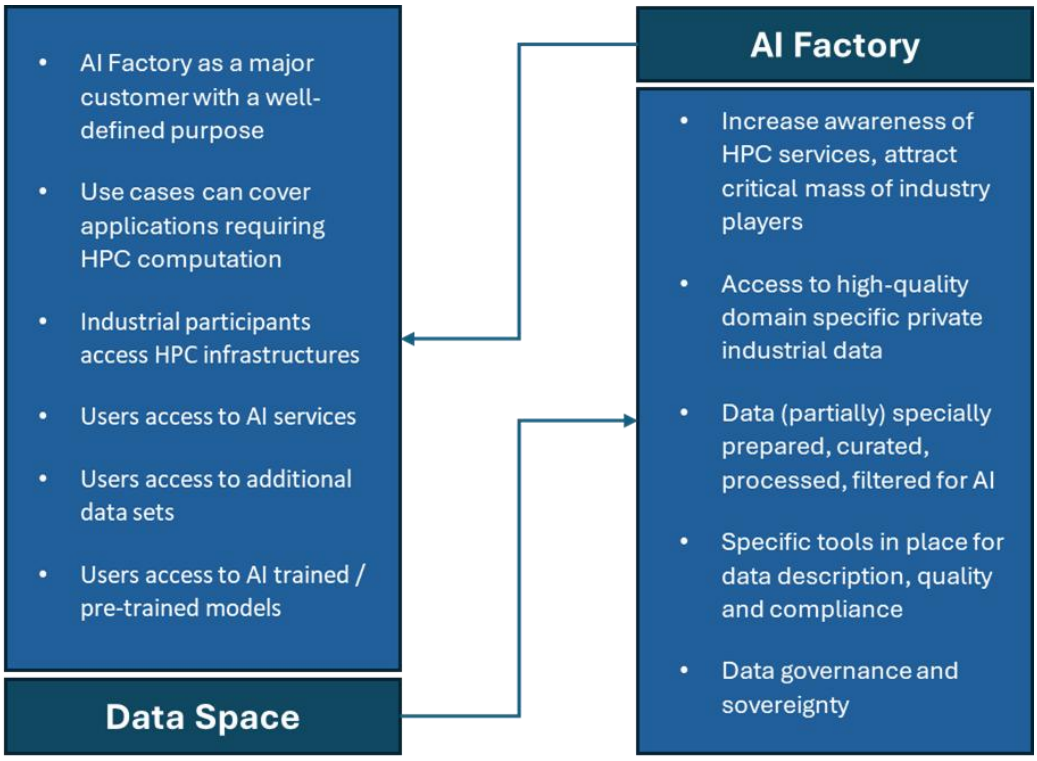

Figure 1. Mutually beneficial relationship between AI Factories and EU Data Spaces

3. Laying the foundations for collaboration

3.1. Collaboration models and scenarios

The way the AI Factory and data space collaborate and interact fully determines the level of integration, challenges, and the value generated. Thus, the ultimate goal would be to select a collaboration approach that maximizes value while minimizing effort, something quite difficult to achieve at this stage. Therefore, the purpose of this section is just to outline some different scenarios to be further explored in subsequent versions of the paper, always considering that this is a key aspect in the discussion. Notice that an AIF can be participant of many data spaces, where benefit can be mutual.

The list of scenarios below increases the level of integration (see Figure 2):

- **AI Factory and data spaces as separate entities collaborating in an "ad-hoc" fashion**, based on a flexible, informal partnership where each entity maintains its independence, and collaboration happens only when specific needs or opportunities arise. The collaboration between the AI Factory and the data space is initiated as needed, driven by specific projects, use cases, or requests.
- **Collaboration via intermediaries.** In the early stages, the role of intermediaries who support both parties in establishing the relationship and carrying out some crucial tasks will be key. These intermediaries would support the AI Factory to find and contact the most appropriate data space, according to its interest (in a current scenario where there is not still a single entry point to access all EU Data Spaces, and where DSSC can help with its holistic view of EU landscape). They will also support AI Factories in developing critical capabilities such as data connection, trusted data transactions, and governance frameworks. They will also provide intermediary services, that eventually could contribute to support interoperability with other AI factories and data spaces.

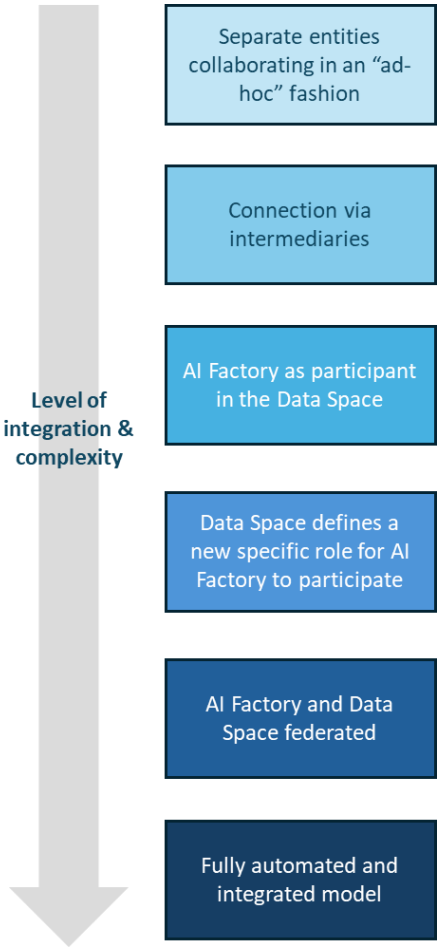

Figure 2. Level of integration between the AI Factory and the EU Data Space

- **AI Factory as (any other) participant in the Data Space**, where the AI Factory is treated like any other participant in the data space, following the same rules, governance, and protocols. In this model, the AI Factory becomes an active member of the data space, just like any other organization (e.g., companies, SMEs, research institutions) that contributes to or consumes data. The AI Factory follows the same rules and governance structures as other participants in the ecosystem.

- **The Data Space defines a new specific role for the AI Factory to participate in the data space**. In this scenario, the AI Factory is recognized as a new, specific role within the data space, with unique responsibilities and privileges tailored to its function. Unlike merely being another participant, the AI Factory is acknowledged as playing a distinctive and crucial role in the data space ecosystem, with an additional level of trust, so (some) data can be brought to the AIF and locally mirrored more easily.
- **AI Factory federated with the data space**, where they operate in a collaborative relationship while maintaining a degree of autonomy. This approach allows for integrated functionalities and shared resources, but each entity retains its individual identity and operational control.
- **Fully automated and integrated model**, where they operate as a seamlessly connected system, leveraging technology to enable real-time data exchange, resource sharing, and automated workflows. This model aims to streamline processes and minimize manual intervention, resulting in enhanced efficiency and speed of operations.

3.2. Guidelines (how to start: from lightweight to a closer interaction)

As previously noted, the connection between AI Factories and Data Spaces offers a promising win-win scenario worth pursuing. However, establishing this connection is a complex undertaking, as it involves integrating two intricate ecosystems, each with distinct objectives, requirements, governance structures, and stakeholders. Additionally, there is not much time to establish the connection. Finally, many European Data Spaces are still under development, making it difficult to propose specific and detailed actions that AI Factories can take to effectively connect with them at this stage. Notice also that the process would depend on the type of collaboration scenario selected, as explained in the previous section.

However, leveraging the existing knowledge and resources from initiatives such as the Data Space Support Center (in particular DSSC co-creation method[9]), we can outline a set of fundamental steps that are likely to apply across most potential scenarios, and that it can be considered as a ***"lightweight" approach for AI Factories to connect with data spaces***.

Contact **data space governance authority.** Consider initially for the AI Factory to become a data space participant. This initial contact is crucial for the AI Factory to understand the requirements and expectations for becoming a participant in the data space. Engaging with the governance authority will also provide insights into the onboarding process and any necessary documentation.

1. Review **data space governance framework** (rulebook). The rulebook would include, among others, policies at data space level, data models used in the data space and (type of) data driven services available. If the framework aligns with AI Factory's goals, ensure adherence to these general policies and start the onboarding process.
2. Check **mechanisms in place for data exchange and retrieval.** Investigate the mechanisms available for exchanging or retrieving data from the data space (include identifying the data space participant agent or connector that enables access) and familiarize with the protocols and APIs that govern data interactions, ensuring they are compatible with the AI Factory's systems.

[9] https://dssc.eu/space/bv15e/766062883/Co-Creation+Method

3. Consider the **optimal connectivity** between data space and AIF to carry out the above exchange of data. Assess the optimal connectivity solutions between the data space and the AI Factory to facilitate efficient data exchange. Consider factors such as bandwidth, latency, and data transfer rates to ensure that the connection can support the expected volume and frequency of data transactions. Evaluate whether direct connections, data pipelines, or intermediary services best meet the needs.
4. Establish **clear business conditions** for participation in the data space. This includes discussing and agreeing on pricing models, such as regular fees, pay-per-use, or revenue-sharing arrangements. It's essential to align these conditions with AI Factory's cost structure and revenue model to ensure a sustainable partnership. Document these agreements to provide a clear framework for ongoing collaboration.

The data space would provide mechanisms and conditions to scale-up the access and reuse of data. Accordingly, a ***second*** stage could review the relationship between the AIF and the data space:

1. (+) Assess whether the relationship between the AI Factory and the data space should evolve beyond mere participation. Consider moving towards a federated scheme or other collaborative mechanisms discussed in the previous section.
2. (+) Check whether it is possible to access and replicate / instantiate (part of) the **(federated) catalogue** of datasets (including metadata descriptions). Include metadata descriptions more tailored to the specific needs of AI practitioners and specific use cases. Data quality properties and metrics tailored to the intended use of the data to train AI models. Focus on the datasets intended to train high-risk AI systems, whose quality description should include, according to AI Act: relevance, representativeness, free of errors, complete, appropriate statistical properties, specific geographical and behavioral / functional setting and appropriate safeguards for the fundamental rights and freedoms of natural persons)
3. (+) Determine the most suitable data exchange mechanisms for connecting the AI Factory with the data space. This includes evaluating specialized connectors and protocols that enhance data interoperability and facilitate seamless data transactions. Review the existing connectivity options to ensure they can accommodate the expected data flow and functionality required for the AI Factory's operations. Consider performance metrics, reliability, and scalability when selecting these mechanisms.
4. (+) Consider **end-to-end** data governance, data protection and cybersecurity
5. (+) Explore **advanced business models** more suitable for an advanced interaction

In any case, and as outlined in the previous section, a key preliminary step for an AI Factory is to identify the data space it wishes to connect with, based on its specific needs and interests. In this process, the **role of an intermediary** is crucial, offering guidance and support (with DSSC being of guidance here). Notice also that trusted intermediaries can support AI Factories and Data Spaces on the different proposed steps.

3.3. Potential business models

In order for the connection between AI Factories and data spaces to succeed, they should be able to agree on joint business models that result in value to all stakeholders involved in both sides.

The Data Spaces Support Centre devotes a whole pillar of its blueprint to business, and a specific building block to data space business model development[10], and some of this content can be useful to start outlining potential joint business models (data driven canvas model, co-creation questions, etc. …). However, this exercise should be extended to consider costs associated with accessing the various components and activities outlined in the left side of Figure 1 (section 2), but also with the potential revenues resulting of them.

Similarly, this exercise should be extended to consider how to integrate into the **business model of the AIF** the **new conditions and costs associated** with the process of accessing, acquiring and reusing data. These costs can range from a general fee to subscribe to the data space (granting access to the entire data catalogue), to pay-per-use, or price to pay for specific data sources and / or data providers, according to initial policies and conditions attached to the datasets, and potentially elaborated further in contracts negotiation. Additionally, the AI Factory must consider the costs associated with the delivery of new data-driven services. Conversely, the AI Factory may generate revenue from the reuse of this data by end users, potentially leading to innovative collaborative business models that manage and distribute the value exchange between AI Factories and data spaces.

Considering all the above, AI Factories and data spaces will have to explore **new and innovative business models** and opportunities that involve stakeholders from both sides, including industrial communities (providing private data) and research and academia, so they can benefit from each other, and eventually lower barriers of access to private data for innovators. And this is a matter of both control and value for companies: if they contribute, they need not only to be sure that they are managing their data well, but they also need to be sure that the data sharing has a value for them in terms of cost, performance or revenue. There are several sources of power in a value chain (e.g. using Porter's five forces model to analyse dynamics[11]) and the AI Factory needs to consider how data can fit these and will affect suppliers, consumers, substitutes, competitors, and new entrants.

Finally, it would be worth exploring the use of open AI models as proof-of-concept within the AI Factory. These models could be hosted for early-phase projects that utilize few-shot prompting techniques. The further development of these open AI models can lead to a variety of AI-based tools designed to intelligently support SMEs by leveraging data and computational resources. Additionally, numerous open-source AI models developed by researchers hold significant potential for integration within the AI Factory. However, gaining a comprehensive overview of these available models can be challenging, highlighting the need for a systematic approach to facilitate access and connection to these valuable resources.

[10] https://dssc.eu/space/bv15e/766064638/Business+Model

[11] https://en.wikipedia.org/wiki/Porter%27s_five_forces_analysis

4. Key considerations: challenges and opportunities

The realization of the integration between AI factories and Data Spaces requires establishing a well-defined roadmap of activities and continuation of investments in the Data Spaces programme. This section outlines some of the key considerations to take into account when addressing this connection.

The first key consideration is the need **to align the timing between AI Factories and Data Spaces**, given their differing levels of maturity. At the time the first calls for AI Factories are published (September 2024) and potential hosting organisations are defining their strategies, the overall EU Data Space deployment programme is underway, setting up the main European infrastructures for secure, sovereign and trustworthy data sharing. pooling user companies, public sector, research and individuals to share data in a secure and trustworthy manner, facilitating trust and data sovereignty. Therefore, Factories will, in many respects, build on the foundational work of Data Spaces. To ensure a **seamless transition and avoid redundancy**, it is crucial to align these efforts. AI Factories can play a key role in accelerating the availability of data within Data Spaces, particularly non-open, copyrighted material needed for training foundation models. Besides, while **AI Factories primarily operate at a national level**, focusing on local strategies and innovation ecosystems, **EU Data Spaces have a broader European dimension**, aimed at fostering cross-border data sharing and collaboration across multiple sectors. Reconciling these two perspectives requires careful coordination to ensure that national AI Factory initiatives are aligned with the overarching goals of the European Data Spaces.

Data availability is at the core of this connection. Data spaces and related core access technologies are currently being gradually established, while the training of large-scale foundation models already requires constantly growing large datasets today. It means that the relevance of data spaces to potential AI Factories users will also only increase gradually, which could affect the early adoption of data space technology within AI Factories. Data spaces should consider providing in time and quality the constantly growing large datasets that the training of large-scale foundation models requires[12], in terms of volume, but also clear policies about security and data management through the whole lifecycle. Data spaces need to be optimised to provide data in a format that supports both pre-training and fine-tuning with lots of tokens and labels. Data should be optimized to fit well within the context windows of typical LLMs, so a hierarchical product approach (component-subcomponents) will be the difference for many applications.

This implies that the **data lifecycle** will be shared between the Data Space and the AI Factory, requiring coordination on when and how various data processing techniques—such as cleaning, quality assessment, and curation—will be applied. For example, data quality will be addressed at multiple stages (provider, Data Space, AI Factory, end user), with each party ensuring quality at their respective stage, all aimed at ensuring the data meets the needs of the final application. Eventually, the AI Factory should guarantee the **quality of data** put at the disposal of users, whose description should consider specific data quality dimensions and metrics for AI, but more importantly, those **reflected in the AI Act** and especially when used **to train high-risk systems.** Additionally**,** the AIF would include mechanism or services to assess the

[12] https://arxiv.org/abs/2203.15556: large AI models needed approx. 20x tokens compared to the number of parameters[12]. Hence, an AI Factory needs to provide 20B tokens to pre-train a model of 1B parameters

quality of the data (according to the quality description) and processes in place to address the **liability** about the declared quality of the data

Traditionally, HPC ecosystems have worked in isolation, connecting just with local data facilities. This is now changing, for example, with the new action to federate HPC CCs, but still **how to connect them to other digital ecosystems** both for computation purposes (federation of HPC centres, connection with cloud, edge devices, … –digital continuum) and with data storage facilities (cold / hot / ephemeral), including data spaces, imposes some challenges. Both approaches require well-defined workflows (here in the context of AI approach) and orchestration, including physical connectivity (considering high-bandwidth connections, dedicated lines, advanced networking technologies such as optical fibers, or high-speed internet backbones), but also connectivity at other levels (security, coordination, federation…).

In this context, it is important to consider the potential **low latency challenge between high-performance computing (HPC) and data spaces**, as this could impact the performance of supercomputers, which are increasingly focused on supporting rapid processing

Once connectivity is set up, **interoperability between the AI Factory and the data spaces** emerges as a key factor to consider. It includes interoperability at different levels[13]. This should be addressed considering existing international standards and upcoming ones and aligned with different ongoing efforts (EC data standardization request on data and data spaces, CEN CENELEC Trusted Data Transactions, CEN CENELEC Focus Group Data, Dataspaces, Cloud and Edge, CEN CENELEC JTC25). The challenge of interoperability is heightened by the fact that AI Factories have to deal currently with the fragmented ecosystems of data spaces, that compel them to approach individually each data space they want to connect to.

There will also be needed a **collaborative scheme** (according to capabilities of the data ecosystem and lifecycle of data and models) to ensure proper **governance of data** about who can access what data and under what conditions (especially considering copyright clearance for some types of data), data sovereignty, the conditions for the reuse of data, data traceability) **and governance of AI**, and compliance assurance (GDPR, DGA, AI Act).

This scheme should ensure **data provenance and traceability**, so that the data used by AI Factory customers to train models for their products remains accessible for years after the initial training. In the event of a product malfunction or accident, legal authorities may require this data to identify potential errors made during the product's development. AI Factory customers must be able to provide this data when requested. It is the responsibility of the AI Factory and data providers to guarantee data availability, or alternatively, the product owner may need to store a copy themselves (e.g. the automotive industry currently retains all data used to train self-driving systems, a practice also relevant to fields like biomedicine and healthcare).

And the collaborative scheme should also provide, on an **end-to-end basis, integrity, inviolability and privacy** of the acquired data, when going through the different processing and computational processes, and when delivered to end users. It has also to be considered that due to the current lack of enough appropriate data in European data spaces, users may process inside AI Factories data that originates from outside the EU, that might raise conflicts with European and particulate data spaces values. AI Factories need to address related risks and take responsibility and establish monitoring and awareness to retain

[13] https://ec.europa.eu/isa2/eif_en/

trust within the ecosystem when transitioning towards fully trustworthy data spaces (**ethical responsibility in transitional data ecosystems).**

AI Factories and data spaces should also agree on a common approach for **identity and role management**, to register and identify users and to determine access rights. This is a challenge for both (i) data from multiple data sources used in one factory, and (ii) data from one source being used in multiple factories. Only a common framework for classification with respect to export control, nationality of the user, place of use of the result will allow seamless access grants.

Additionally, and depending on the collaboration model between the AI Factory and the data space (section 3.1), the AI Factory and the data space should agree on the right mechanisms to enable **the actual exchange of data** from the data spaces (APIs, protocols, connectors, …), guaranteeing the trustworthiness of the transaction. Existing initiatives working on this field (e.g. CEN CENELEC Trusted Data Transactions, Data Space Support Center) should be considered. Based on the above, the AIF should provide, when needed, an environment that **guarantees trusted and secure operations** through the data acquired and processed.

The integration of business models from AI Factory and data space (as discussed in section 3.3) should also allow their **business scalability**, considering broader usages in research and industrial applications, and providing benefits/support from small to big, from established to new businesses. But scalability should also target the **automatization** of the access, retrieval, and reuse of data so that the process can scale-up and AI Factories are able to provide the increasing amount of data required for large AI models. Scalability also includes efforts to bring all important and relevant stakeholders on board and foster the use and adoption with corresponding incentives.

Finally, there is a critical need for **appropriate skills within AI Factories** for the acquisition, management, and governance of data, including the role of a Data Protection Officer. These skills are essential for training stakeholders and supporting end users. Furthermore, depending on the level of abstraction and support offered by AI Factories, end users will also need skills to effectively manage and utilize data in accordance with its intended purpose.

5. Recommendations

- Define a **unified approach for AI Factories and EU Data Spaces**. Develop a collaborative strategy that harmonizes connections between AI Factories and EU Data Spaces. This will help avoid ad-hoc solutions for each AIF connecting to specific data spaces. Establish a high-level alignment on the mechanisms for connecting, accessing, retrieving, and managing data, including a common framework that ensures effective AI and data governance, shared with the data space. This framework should incorporate measures to ensure compliance with data regulations (such as GDPR, DGA, and the AI Act), distributing the responsibility for governance and compliance assurance according to the lifecycle of data and AI models.
- **Realize a federation of EU Data Spaces to provide a single-entry point to the whole EU data ecosystem.** Create a unified access point that addresses the current fragmentation among data spaces, alleviating the burden on AI Factories by eliminating the need for individual connections to each specific data space.
- Leverage on the upcoming HPC Federation **to align, connect, and integrate** AI Factories (HPC CCs) **with existing and emerging digital ecosystems**, particularly data spaces across Europe and internationally. This evolution is essential for supporting European businesses, as access to specific datasets will become the exception, with data retrieval from data spaces becoming more common. This framework will also facilitate the connection with other AI-driven ecosystems with whom they are also expected (as required in the calls) to share data and other contents, especially the European AI on demand platform (through the DeployAI project), but also EDIHs and AI TEFs. This connection should also include other European initiatives like SIMPL, High Level Forum Standardization, Data Space activities from EC and European countries.
- **Continue learning based on real hands-on experiences of the first AI Factories**, enhancing the collaboration between them and ongoing EU Data Spaces. We recommend **BDVA as the ideal forum for these discussions**, given its membership in the EuroHPC JU, partnership in the Data Spaces Support Centre, and founding and active role in Adra.
- To move quickly towards this strategy, initiate the **creation of a Working Group** that includes representatives from the European Commission, EuroHPC JU, AI experts from supercomputing centers, and data spaces utilizing HPC (such as LDS, EUCAIM, and the EU Genomic Data Infrastructure). Additionally, incorporate perspectives from the AI Office, academia, and industry stakeholders. Finally, consider the national AI strategies of each Member State to ensure complete alignment.
- Consider the inclusion of experienced **third parties that can serve as trusted intermediaries,** to provide services and support AI Factories in developing critical capabilities such as data connection, trusted data transactions, and governance frameworks. These intermediaries would provide essential guidance in the short term, ensuring that AI Factories can operate effectively while waiting for data spaces to offer full support.
- Identify **existing standards** that can facilitate the integration of AI Factories with EU Data Spaces, as well as pinpoint any potential gaps in these standards. This process should involve a comprehensive

review of relevant industry standards, best practices, and regulatory requirements that support data interoperability, security, and governance.

- **Operationalize data ethics,** leveraging on the experience of previous initiatives (e.g. EUHubs4Data[14] and its ethics toolkit[15]) or by the establishment of a transparent data management platform (containing published data management plans i.e. via systems like Argos OpenAIRE[16]) at each AI Factory, which addresses specially requirements about privacy and data protection (extension of PET technologies from data spaces[17] to AI Factories). This exercise will allow AI Factories to retain trust when connecting with dataspaces, support compliance with current and future AI Act regulations on data provenance for foundational models and demonstrate the successful use of data sets from European Data Spaces. This would also involve the integration of data and AI risk management inside AIFs and offer appropriate training to users.
- Establish **demonstrator or lighthouse projects** showcasing practical implementations of AI Factory connections to Data Spaces. These projects will serve as concrete examples to accelerate the shared understanding of integration processes and operational frameworks among stakeholders
- Explore **new and innovative business models** and opportunities between involving all stakeholders from both sides, so they can benefit from each other, and eventually lower barriers of access to private data for innovators.

[14] https://euhubs4data.eu/
[15] https://github.com/EUH4DEthics/OpenCallDataInnovationEthicsToolkit
[16] https://argos.openaire.eu/splash/
[17] https://bdva.eu/download/92/publications/5196/leveraging-the-benefits-of-combining-data-spaces-and-privacy-enhancing-technologies.pdf

Annex. Activities and contributors

This document reflects the outcomes of a series of discussions, meetings, sessions, and activities organized by the BDVA with different members of the community and stakeholders.

These efforts are part of a strategic plan established following the workshop held by the European Commission on May 17, 2024, aimed at supporting the successful development of AI Factories, particularly their integration with EU Data Spaces. This includes:

- Several meetings with policy makers
- Dedicated sessions with experts in BDAV Task Force HPC/Big Data/AI
- Meetings with some HPC CCs potentially hosting an AI Factory
- Meetings with some EU Data Spaces
- Dedicated session with the BDVA Technical Board
- Dedicated sessions at the European Big Data Value Forum 2024: “AI Factories: Addressing the data challenge” (https://european-big-data-value-forum.eu/session/ai-factories-addressing-the-data-challenge/).

The list of the contributors is the following (alphabetical order):

- Aleksi Kallio (CSC-IT)
- Ana Garcia (BDVA)
- Ana Isabel Torre Bastida (Tecnalia)
- Andrejs Vasiļjevs (Tilde)
- Daniel Alonso (BDVA)
- Damien Lecarpentier (CSC-IT)
- Ed Curry (Insight/University of Galway, BDVA Vp)
- Georg Rehm (DFKI)
- Jansik Branislav (IT4Innovations National Supercomputing Center)
- Jeanette Nilsson (RISE)
- Katerina Slaninova (IT4Innovations National Supercomputing Center)
- Laura Morselli (CINECA)
- Maria Jose Lopez Osa (Tecnalia)
- Maria Perez (UPM)
- Mike Matton (VRT)
- Laure Le Bars (SAP, BDVA Vp)
- Roberta Turra (CINECA)
- Sara Garavelli (CSC-IT)
- Savvas Rogotis (BDVA)
- Sergi Girona (BSC)
- Stelios Piperidis (Athena)
- Thomas Hahn (Siemens, BDVA President)
- Till Riedel (KIT)
- Tuomo Tuikka (VTT)
- Ulrich Thombansen (Fraunhofer)

Interested in contributing? Contact us at: innovation@bdva.eu