

GENERATIVE AI AND DATA SPACES

DSSC White Paper



DATA SPACES
SUPPORT CENTRE

Publisher

[Data Spaces Support Centre \(DSSC\)](#)

c/o Fraunhofer-Gesellschaft zur Förderung
der angewandten Forschung e. V.
Hansastr. 27c
80686 Munich
Germany

Copyright

Data Spaces Support Centre, September 2024

Consortium

[Big Data Value Association](#)

[CapGemini Invent](#)

[FIWARE Foundation](#)

[Fraunhofer-Gesellschaft](#)

[Gaia-X](#)

[International Data Spaces Association](#)

[KU Leuven](#)

[MyData Global](#)

[SITRA](#)

[TNO](#)

[University of Galway](#)

[VTT](#)



**Funded by
the European Union**

Contact

www.dssc.eu

contact@dssc.eu

The Data Spaces Support Center receives funding from the European Union Digital Europe Programme under grant agreement n° 101083412.



Data Spaces Support Centre; October 2025; version 1.1*; CC BY-SA 4.0

* Read the change log at page 127 to observe the modifications compared to v1.0

Editors

Ana Garcia, Big Data Value Association (BDVA)

Savvas Rogotis, Big Data Value Association (BDVA)

Eimear Farrell, European Commission, Joint Research Centre (JRC), Ispra, Italy

Tobias Guggenberger, Fraunhofer Institute for Software and Systems Engineering (Fraunhofer ISST)

List of Contributors

Sincere gratitude needs to be expressed to all those who contributed to the development and production of this second collaborative paper of the Data Spaces Support Centre and its Strategic Stakeholder Forum, on the topic of "Generative AI and Data Spaces". The creation of this paper has been made possible through the collective efforts of an extensive number of contributors representing very relevant and complementary organisations and initiatives on the focus topic. Appreciation is extended to everyone who took part in the discussion meetings that were organised, provided feedback directly in the early versions of the paper and shared their valuable insights. All the inputs received have significantly shaped the final outcome.

The list of contributors (individuals, organisations & initiatives) is provided below (in alphabetical order):

- **From DSSC Consortium partners:** Daniel Alonso (BDVA), Ana Garcia Robles (BDVA), Matteo Frigeri (KUL), Tobias Guggenberger (Fraunhofer ISST), Arash Hajikhani (VTT), Rafiqul Haque (Insight Centre for Data Analytics, also connecting the Clarify project), Atte Kinnula (VTT), Leona King (KUL), Marko Komssi (VTT), Thomas Margoni (KUL), Daham Mohammed Mustafa (Fraunhofer FIT), Satu Salminen (SITRA), Savvas Rogotis (BDVA), Tuomo Tuikka (VTT).
- **From Common European Data Spaces:**
 - Green Deal Data Space (GREAT): Gabriele Accarino (CMCC), Donatello Elia (CMCC), Marta Gutierrez (EGI), Marco Parrillo (EGI), Mattia Santoro (CNR).
 - Language Data Space (LDS): Khalid Choukri (ELDA), Stelios Piperidis (Athena Research Centre), Georg Rehm (DFKI), Andrejs Vasiljevs (TILDE also in his capacity as GenAI Task Force lead in BDVA).
 - Trusted Media Data Space (TEMS): Emeline Banzuzi (1001 Lakes), Marko Turpeinen (1001 Lakes - also included in the discussions as SSF member)

- **From other contributing organisations that are part of the DSSC Strategic Stakeholder Forum:** Kuldar Aas (Ministry of Economic Affairs and Communications of Estonia), Eimear Farrell (European Commission, Joint Research Centre (JRC), Ispra, Italy), Sara Garavelli (CSC), Elisa Halonen (CSC), Gerard van der Hoeven (iShare), Panos Ilias (ILVO - also connecting the Soilwise project), Aleksi Kallio (CSC), Alexander Kotsev (European Commission, Joint Research Centre (JRC), Ispra, Italy), David Fernandez Llorca (European Commission, Joint Research Centre (JRC), Seville, Spain), Marco Minghini (European Commission, Joint Research Centre (JRC), Ispra, Italy), Ignacio Sanchez (European Commission, Joint Research Centre (JRC), Ispra, Italy), Serena Signorelli (European Commission, Joint Research Centre (JRC), Ispra, Italy).
- **From other relevant AI initiatives part of the Network of Stakeholders (NoS) of DSSC:**
 - AI Testing and Experimentation Facilities:
 - agrifoodTEF: Fabio Pianesi (FBK) - also representing Dynamo Cloud (SSF member)
 - AI-MATTERS: Valentina Ivanova (CEA), Christophe Leroux (CEA)
 - TEF-HEALTH: Petra Ritter (BIH-Charite)
 - AI on Demand Platform (DeployAI): Denia Kanellopoulou (Demokritos), Joachim Köhler (Fraunhofer IAIS also connecting OpenGPT-X project), Johannes Leveling (Fraunhofer IAIS also connecting OpenGPT-X project).

The paper has been peer-reviewed by Clara Pezuela (FIWARE).

Designed and proofread by Daniel Djamo (BDVA).



Executive Summary

The rise of Generative AI (GenAI), underpinned by new foundation models, offers numerous opportunities for businesses and society in Europe.[1] Data spaces are interoperable frameworks that enable data transactions between participants. Based on common governance principles, standards, practices and enabling services, they offer unprecedented possibilities to unlock the data economy. We believe that the synergy between GenAI and data spaces hold significant transformative potential. It can be one of the keys to unlocking the value of industrial, commercial and personal data for Europe and worldwide. These synergies can enable the use of data in an ethical, trustworthy and regulatory-compliant manner, allowing for fairness in data and GenAI value chains, as well as creating new business opportunities and incentives for data sharing.

GenAI developments face several challenges in Europe and concrete needs are becoming apparent throughout the entire lifecycle - from model creation to deployment and usage. These needs go beyond data-related necessities. They encompass computing power needs, legal and ethical requirements and the need for specialised skills. Other factors like business and application requirements, financial resources, time constraints, value creation and environmental impact are also very relevant in this discussion. Understanding these multidimensional challenges is crucial for unlocking the full potential of GenAI in Europe.

Data spaces have already begun to address some of these needs and through significant EU investments to open new opportunities. They provide secure and efficient frameworks for data sharing and enable access to diverse, domain-specific datasets - including proprietary industrial data - which are critical for GenAI developments. Data spaces offer strong foundations for data quality and relevance, key aspects for trustworthy and ethical AI. Beyond data sharing, data spaces bring up an ecosystem perspective, fostering innovation and collaboration among involved stakeholders which can create value, form partnerships and generate new business opportunities. They also promote compliance with relevant EU regulations like the General Data Protection Regulation and the Artificial Intelligence Act, thus, ensuring transparency, accountability, and adherence to legal obligations. Moreover, through the use of standards, data spaces allow for interoperability and cross-domain synergies, while paving the way to meet data governance, intellectual property, and cybersecurity needs.

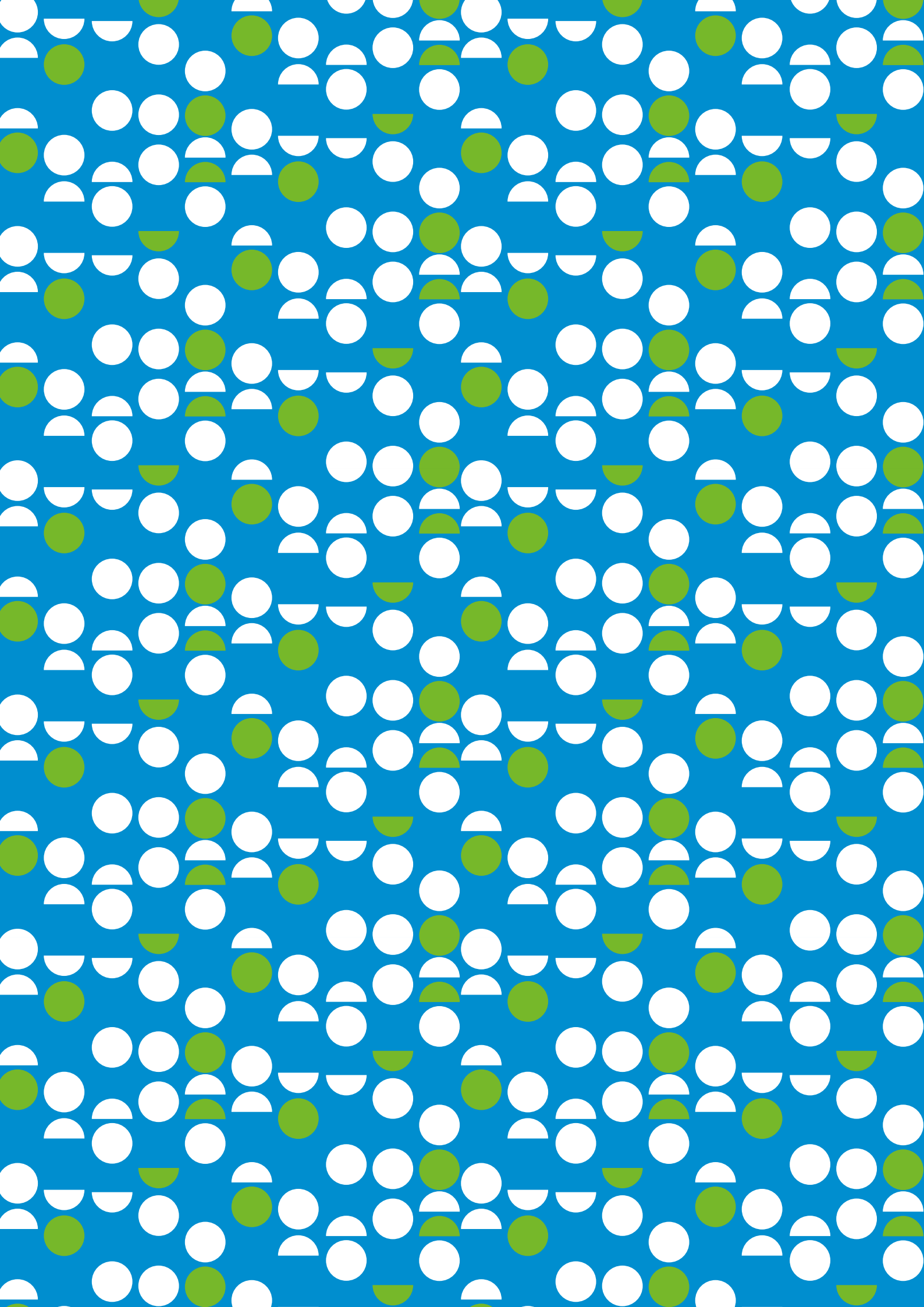
[1] <https://bdva.eu/task-forces/generative-ai/>

However, this interplay come with a specific set of limitations. In order to harness the full potential that data spaces can offer to GenAI we still need to address several key challenges. Indicatively, data spaces must evolve and become more mature to meet the increasing demands of GenAI. The latter, is followed by the need for a holistic framework and integrated digital infrastructure that can support GenAI on a larger scale. Data spaces under their current setup need to improve in terms of interoperability and simplicity, while ensuring diversity, quality, and representativity in data is another key priority.

In line with the above-mentioned challenges, a series of 8 recommendations is elicited and is oriented towards different target audiences. Some of the recommendations are targeting towards data space initiatives and the DSSC (for instance: define and deploy means of simplified, harmonised access to cross-sectoral and federated data spaces serving data accessibility for GenAI needs – identify use cases and to create of a catalogue of learnings from GenAI applications built on top of data sharing) and other recommendations are more transversal, referring to a broader set of stakeholders including policymakers, industry and existing industrial ecosystems; European GenAI infrastructure initiatives and to a lesser extend GenAI providers (for instance: coordinate efforts for creating domain-specific GenAI models across all industrial sectors by building on common European data spaces – continue balanced and coordinated investments in data spaces, GenAI, and computing infrastructures).

Several common European data spaces and GenAI projects already offer promising learnings and illustrations of the value of data spaces for GenAI. Additionally, research projects across various domains supply valuable insights and experiences.

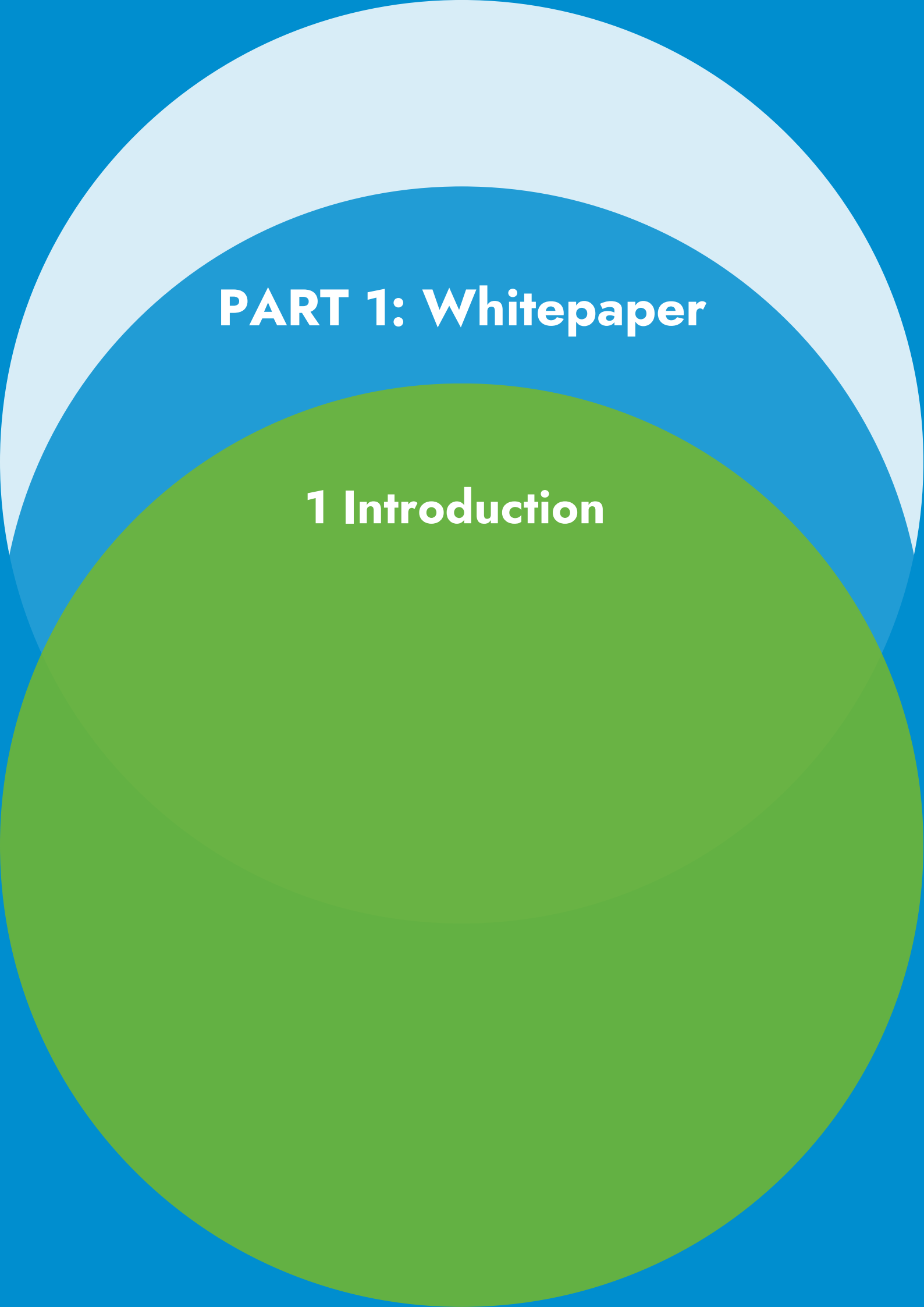
The whitepaper also presents AI Factories as a case study. Introduced by the EC as part of its AI Innovation Package,[2] the AI Factories rely on the combination of data, computing power, and talent to promote GenAI advancements in Europe and support AI startups, SMEs, and research groups through relevant resources. Integrating AI Factories with common European data spaces can serve to enhance data access and interoperability. However, challenges like data quality, governance, and scalability should be addressed to guarantee a smooth access by AI Factories to this data. A well-defined roadmap is needed to align AI Factories with data spaces and develop innovative business models.



<u>Publisher</u>	2
<u>Copyright</u>	2
<u>Contact</u>	2
<u>Consortium</u>	2
<u>Contributors</u>	4
<u>Executive Summary</u>	6
 <u>PART 1: Whitepaper</u>	 13
 <u>1 Introduction</u>	 13
<u>1.1 Context, purpose and motivation</u>	14
<u>1.2 Generative AI: an opportunity for Europe</u>	19
<u>2 A multidimensional perspective of the challenges of Generative AI</u>	21
<u>2.1 Data challenges</u>	27
<u>2.2 Computing needs</u>	29
<u>2.3 Skills</u>	30
<u>2.4 Legal aspects</u>	32
<u>2.4.1 The EU AI Act - a comprehensive approach to regulating AI</u>	33
<u>2.4.2 Data Governance regulation</u>	37
<u>2.4.3 Regulation of Intellectual Property Rights (IPR)</u>	40
<u>2.4.4 Privacy and Data Protection regulation</u>	43
<u>2.4.5 Cybersecurity regulation</u>	46
<u>2.5 Business and financial sustainability aspects</u>	47
<u>2.6 Some other relevant challenges</u>	48
<u>3 The transformative potential of data spaces for Generative AI</u>	52
<u>4 Challenges to unlock the full potential of data spaces for Generative AI</u>	62
<u>5 The specific case study of AI Factories</u>	67
<u>6 Recommendations</u>	77
<u>7 Conclusions</u>	80

<u>PART 2: Learnings from several initiatives</u>	80
<u>1 Learnings from some common European data space initiatives</u>	82
<u>1.1 Media domain: Trusted European Media Data Space (TEMS)</u>	83
<u>1.2 Language Data: Common European Language Data Space (LDS)</u>	85
<u>1.3 Green Deal domain: Common European Green Deal Data Space (GREAT)</u>	87
<u>2 Learnings from some Generative AI projects</u>	90
<u>2.1 Language Data: OpenGPT-X case</u>	91
<u>2.2 Public Tenders Data: Tenders Electronic Daily (TED)</u>	94
<u>3 Learnings from several sectors based on research activities and projects</u>	96
<u>3.1 Healthcare domain: Clarify project case</u>	97
<u>3.2 Cybersecurity domain: Addressing Data Availability Issues in Mass Markets</u>	99
<u>3.3 Agriculture domain: SoilWise project</u>	101
 <u>Annex A EU policy framework and relevant initiatives</u>	 103
<u>A.1 The EU vision for a data-agile economy</u>	105
<u>A.2 EU AI-related policy landscape</u>	107
<u>A.3 The role of harmonised standards in policy implementation</u>	112
<u>A.4 Relevant European actions (projects & programmes)</u>	114
<u>A.4.1 HPC/Computing community</u>	114
<u>A.4.2 AI on Demand Platform (AIoDP)</u>	115
<u>A.4.3 AI Testing and Experimentation Facilities</u>	116
<u>A.4.4 AI Factories</u>	117
<u>A.4.5 Other relevant initiatives in the larger digital ecosystem</u>	117
 <u>Annex B Technological and Industrial trends</u>	 118
<u>B.1 Technology trends and enablers for GenAI</u>	119
<u>B.2 Industry trends for GenAI</u>	121
 <u>Generative AI and data spaces change log</u>	 126

<u>Figure 1 – Symbiotic collaboration between data spaces and GenAI</u>	16
<u>Figure 2 – Challenges along the GenAI project lifecycle</u>	23
<u>Figure 3 – LLMs and multiagent systems in support of automation of workflows</u>	89
<u>Figure 4 – GenAI ‘High Impact Accelerators’ including industrial sectors which are poised to benefit the most from GenAI adoption</u>	124
<u>Figure 5 – GenAI ‘Aspirants’ - industries that while still capable of benefiting from GenAI adoption, they do so to a lesser extent</u>	124
<u>Figure 6 – GenAI ‘Low Potentials’ – industries which that are expected to gain the lowest benefits due high human labour and/or technology legacy</u>	125
<u>Figure 7 – GenAI adoption at country level</u>	125



PART 1: Whitepaper

1 Introduction

1.1 Context, purpose and motivation

Artificial Intelligence (AI) has made considerable progress in recent years, addressing needs and offering solutions to real-world problems and enhancing capabilities across various sectors that impact our daily lives. Advancements in both computing capacity and in data availability and accessibility have dramatically accelerated AI development, including the emergence of Generative AI (GenAI) models and applications. The rise in GenAI, underpinned by new foundation models, offers numerous opportunities for businesses and society in Europe.[3] Within the EU Artificial Intelligence Act (AI Act), GenAI is considered as a type of general-purpose AI model (GPAI),[4] and is described as a highly capable model trained on large and diverse datasets and designed to perform a wide variety of tasks with a significant generality, allowing them to be integrated in a variety of downstream systems.[5]

Data spaces have emerged in the last decade, and thanks to the public and private investments at national and European level, and to the work developed by several European and national associations, we have today access to market projects, demonstrators and success stories that already show purpose and the value created. [6] A data space is “an Interoperable framework, based on common governance principles, standards, practices and enabling services, that enables trusted data transactions between participants. Data space enabling services are implemented by one or more infrastructures”.

[3] <https://bdva.eu/task-forces/generative-ai/>

[4] AI Act recital 99.

[5] For the legal definition of ‘General-Purpose AI model’ see AI Act Art 3 (65). Also see discussion in Fernández-Llorca, D., Gómez, E., Sánchez, I., & Mazzini, G. (2024). An interdisciplinary account of the terminological choices by EU policymakers ahead of the final agreement on the AI Act: AI system, general purpose AI system, foundation model, and generative AI. *Artificial Intelligence and Law*, 1-14.

[6] Why data spaces? A business and user’s perspective. DSSC, 2023.

The common European data spaces were introduced in the EU data strategy[7] and are also referenced in the Data Governance Act[8] and the Data Act,[9] fostering qualities or characteristics related to Interoperability, collaboration and the creation of synergies among data spaces, as well as promoting European values and rules, and the extended purpose beyond business value.

Both GenAI and common European data spaces are interrelated components of the EU digital strategy to foster a robust and secure digital economy and promote innovation.

This whitepaper explores how common European data spaces can play a pivotal role in supporting and accelerating the growth of trustworthy GenAI. We believe the synergy between GenAI and data spaces holds significant transformative potential, for Europe and worldwide, and can be one of the keys to unlocking the value of industrial, commercial and personal data in an ethical and trustworthy manner, allowing fair data in GenAI value chains and creating new business opportunities and new incentives for those sharing the data.

Just as, on the one hand, data spaces can help accelerate the development of European GenAI models and systems, on the other hand GenAI can also be an instrumental tool within data spaces, serving to assist with, and speed up many tasks, thereby also facilitating the adoption. For example, GenAI can be used to help extract value from data, with foundation models/Large Language Models (LLMs)[10] reinforcing and speeding up data analysis within a data space and leading to faster innovation cycles. GenAI can also be used to address data management needs and to automate data lifecycle and data governance processes. In addition, GenAI can allow for the creation of synthetic datasets for exploitation within data spaces. Furthermore, GenAI can support semantic interoperability of data models and APIs, helping to enable FAIR data spaces.[11]

[7] European Commission (2020), Communication on A European strategy for data, COM(2020) 66 final.

[8] Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724.

[9] Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828.

[10] Large Language Models or LLMs are specific subset of foundation models specifically designed for processing and generating human language.

[11] Arnold, B. T., Theissen-Lipp, J., Collarana, D., Lange, C., Geisler, S., Curry, E., & Decker, S. (2024). Towards Enabling FAIR Dataspace Using Large Language Models. arXiv preprint arXiv:2403.15451. <https://arxiv.org/abs/2403.15451>

LLMs can play a valuable role in understanding and translating between different data standards or models, helping to integrate data from various sources within data spaces.

While acknowledging this symbiotic collaboration between Data Spaces and GenAI (visualised in Figure 1), **we have scoped this paper to focus primarily/solely on aspects relating to the potential of Data Spaces for GenAI, in order to help focus the discussion.** Additional aspects (e.g., the possibilities of GenAI for data spaces) may be explored in upcoming papers.

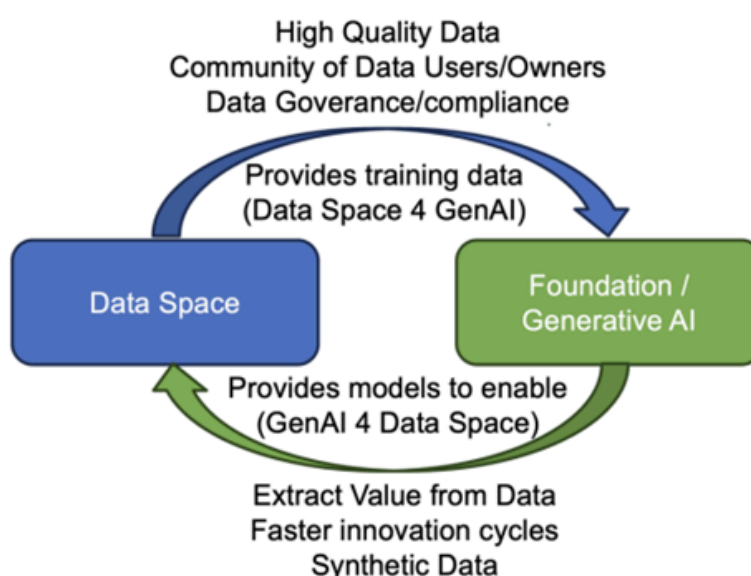


Figure 1 Symbiotic collaboration between data spaces and GenAI[12]

The main goal of this paper is to help inform and enable policy-makers, businesses, data space initiatives, AI initiatives and Gen AI providers (both model providers or GenAI systems providers) to seize the opportunities presented by common European data spaces for supporting and accelerating the growth of trustworthy GenAI. Another important objective of this paper is to foster a common understanding of this topic among various stakeholders and initiatives in Europe, a goal that has been initially achieved during the production of the paper, which as the paper reflects the inputs and knowledge of many different initiatives, experts and professionals.

[12] Curry, E., Zaarour, T., Yang, Y., Timilsina, M., Al-Qatf, M., & Haque, R. (2023, December). Foundation Data Space Models: Bridging the Artificial Intelligence and Data Ecosystems (Vision Paper). In 2023 IEEE International Conference on Big Data (BigData) (pp. 190-195). IEEE.

This whitepaper has been produced in the context of the Data Spaces Support Centre (DSSC)[13] and its Network of Stakeholders (NoS), in particular in collaboration with the Strategic Stakeholder Forum (SSF)[14] but also with several data spaces initiatives that are part of the Community of Practice (CoP) and several AI initiatives that are developing key infrastructures for Europe.

The Data Spaces Support Centre (DSSC) is a key implementation action of the European strategy for data[15] delivered through the Digital Europe Programme (DEP). One of the objectives of the DSSC is to accelerate the deployment, scaling and adoption of data spaces in Europe (and beyond). By fostering the development of the common European data spaces, the DSSC contributes to realising the vision of a genuine European single market for data. The DSSC Strategic Stakeholder Forum (SSF) is a think and-do-tank that supports the DSSC in delivering its objectives. It has an active advisory and strategic role for the DSSC. The organisations that form part of the SSF have expertise on the topic and act as enablers, promoters, or users of data spaces. Some of the organisations in the SSF are also leading or contributing to setting up and deploying data spaces.

In terms of structure and following the introduction (chapter 1), the paper addresses first the needs, requirements and challenges in the different stages or whole value chain of GenAI (what we call the problem statement in chapter 2), and then explores how data spaces help to meet some of those needs and requirements, lower barriers for challenges and unlock additional opportunities (chapter 3). But there are further challenges ahead to unlock this full potential GenAI-Data spaces, and those are described in chapter 4, with a list of recommendations addressed to different target audiences in chapter 6. There is a particular section of this paper (chapter 5) dedicated to the specific case of AI Factories, due to the political and stakeholder attention to this topic, at the moment of releasing this paper. To further complement the discussion from a practical perspective, Part 2 of the paper presents concrete examples and use cases from various domains, illustrating in detail the needs and challenges of GenAI and the potential solutions offered by data spaces. These use cases and examples are sourced from data spaces, GenAI models and applications, as well as other relevant projects.

[13] <https://dssc.eu/>

[14] <https://dssc.eu/space/DC/28049509/Strategic+Stakeholder+Forum>

[15] https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_en

As this paper addresses a large variety of audiences from many different backgrounds this paper incorporates a series of relevant and informative annexes, of mandatory reading for some audiences to understand the flow and content of the paper:

- Annex A: EU policy framework and relevant initiatives. A must-read section for audiences not fully familiar with the European policy framework linked to GenAI and data spaces and the relevant European initiatives.
- Annex B: Technological and industry trends in GenAI

1.2 Generative AI: an opportunity for Europe

Europe's unique blend of diverse languages and cultures, abundant talent in the area of science and technology, and established strengths in the manufacturing and services sectors makes the transformative potential of GenAI particularly compelling across various industries and for society as a whole.[16] Some potential opportunities of GenAI predicted for Europe include - but are not limited to - the following:

- Market analysis indicates that, by 2030, GenAI could potentially boost annual GDP growth in Europe by 0.4% to 0.7%. This represents a significant increase, especially considering that growth without AI effects is projected to be approximately 1.4%.[17]
- The European AI and GenAI market is projected to reach almost \$47.6 billion in 2024 and is forecast to record a compound annual growth rate (CAGR) of 33.7% over the 2022-2027 period.[18]
- The era of AI, particularly GenAI, is just beginning, with OpenAI's GPT-4 API model now available to millions of developers. It is estimated that GenAI could add \$2.6 trillion to \$4.4 trillion annually in global productivity gains, reshaping industries and business models, with banking, high-tech, retail, consumer, and life sciences likely seeing the most substantial impact relative to their revenue. [16]

The following are further examples and perspectives on how GenAI could potentially transform various industrial sectors:

[16] McKinsey & Company. (2023, October 17). Leveraging Generative AI in Europe: The opportunities and challenges. McKinsey & Company. <https://www.mckinsey.com/featured-insights/lifting-europes-ambition/leveraging-generative-ai-in-europe-the-opportunities-and-challenges>

[17] Schlemer, M., Wackerbeck, P., and Ettlin, D. Embracing the GenAI opportunity. How Europe can seize the vast potential of Generative Artificial Intelligence. <https://www.strategyand.pwc.com/de/en/industries/financial-services/embracing-the-genai-opportunity/strategyand-embracing-the-genai-opportunity.pdf>

[18] IDC, Spending on GenAI Solutions in Europe Will Exceed \$30 Billion in 2027, Driving the Overall European AI Market, 21 March 2024, <https://www.idc.com/getdoc.jsp?containerId=prEUR251966524>

- In the financial services sector, several reports claim that by leveraging GenAI, European banks and financial institutions can enhance their services, improve operational efficiency, and strengthen compliance with EU regulations such as GDPR.[16] This not only helps in mitigating risks and preventing fraud but also ensures better customer data protection and privacy. Furthermore, GenAI can support the EU's goals of financial inclusion and economic growth by providing personalised financial services and improving access to banking for underserved populations.
- It is also expected that GenAI will significantly enhance customer experiences in Europe by providing personalised interactions, efficient 24/7 customer support, and intuitive digital interfaces. It can anticipate customer needs, offer proactive solutions, and improve security through real-time fraud detection and compliance with data privacy regulations.
- Furthermore, GenAI is poised to enhance the operations value chain by optimising product development through faster design and prototyping, improving supply chain management with accurate demand forecasting and logistics optimisation, and streamlining manufacturing with process optimisation and real-time quality control. In marketing and sales, AI provides deeper customer insights and better sales forecasting.
- GenAI also has the possibility to enhance public value through improving the efficiency, effectiveness, and reach of both public and social sector initiatives in Europe (for example in the areas of health, environmental protection or education), potentially leading to improved societal outcomes, better public administration and greater quality of life for citizens. [19]
- Based on the above analyses and predictions, it is clear that GenAI has the potential to transform industries and accelerate the evolution of EU companies towards data-driven practices by enabling advanced analytics, predictive capabilities, and operational efficiencies. It will facilitate personalised customer experiences, enhance compliance and risk management, and fosters innovation, thereby boosting competitiveness in global markets. This shift empowers businesses to leverage data effectively for strategic decision-making, operational optimisation, and sustainable growth across diverse sectors in Europe.

[19] European Commission, Joint Research Centre, Tangi, L., Combetto, M., Hupont Torres, I., Farrell, E. and Schade, S., The potential of generative AI in the public sector: Current use, key questions and policy considerations, European Commission, Ispra, 2024, JRC139825. Available at <https://publications.jrc.ec.europa.eu/repository/handle/JRC139825>

2 A multidimensional perspective of the challenges of Generative AI



2. A multidimensional perspective of the challenges of Generative AI

This section focuses on outlining the “problem statement”, in other words describing the main barriers and challenges for GenAI to flourish in the European context. **It considers the whole lifecycle of GenAI, covering the different stages from the model selection or generation to the application deployment and usage.** Those stages are not necessarily linear and consecutive, nor managed or implemented by the same players.

In a paper about GenAI and data spaces one may think that the only relevant GenAI challenges to address are those relating to data. However, addressing other aspects is also important to understand the value of data spaces: e.g., data spaces can provide access to industrial data, but can also lower barriers for some GenAI legal and ethical requirements. Adopting a multidimensional perspective (beyond data needs) helps to identify additional challenges to be addressed to unlock the value of data spaces for GenAI: e.g., the seamless access to high performance computing capabilities from data spaces. So, in addition to data challenges, computing, skills and legal aspects are also considered. All these elements are important in all stages of any GenAI project life cycle, whether this concerns a GenAI model generation, a GenAI system generation, etc. - however the needs and requirements are different.

Data, computing, skills, and legal requirements apply to all stages, but the needs and challenges are different in each stage. Other aspects such as **funding or financial resources, time, specific application needs, and business and value creation** aspects, as well as **environmental impact** are also equally important, however these are not directly addressed (or at least not in depth) in this paper. Algorithm related challenges such as model efficiency, explainability [20], optimisation, scalability, etc are beyond the scope of this paper.

[20] Schneider, Johannes. "Explainable generative ai (genxai): A survey, conceptualization, and research agenda." *Artificial Intelligence Review* 57.11 (2024): 289.

A GenAI project lifecycle incorporates a typical set of phases or stages that need to be followed to achieve defined objectives aligned with specific business/scientific/societal needs. While these stages and their sequence might vary depending on the context of the supported use case, in principle a GenAI application lifecycle includes: Scoping and ideation, model selection or creation, adaptation and fine-tuning; and model evaluation before its final application integration and deployment, and usage. Continuous evaluation and evolution/improvement of deployed models and applications is also needed. It is important to highlight that the different stages can be implemented by different players. The majority of the GenAI lifecycle models add one or more stages for the **data investigation, data acquisition and data preparation**. This is what we call the **data needs or requirements**, that we add as more horizontal stages as is necessary and meaningful during different stages.

Figure 2 provides a visualisation of the different stages of GenAI, as well as the horizontal aspects to consider in the different stages bringing requirements and needs.

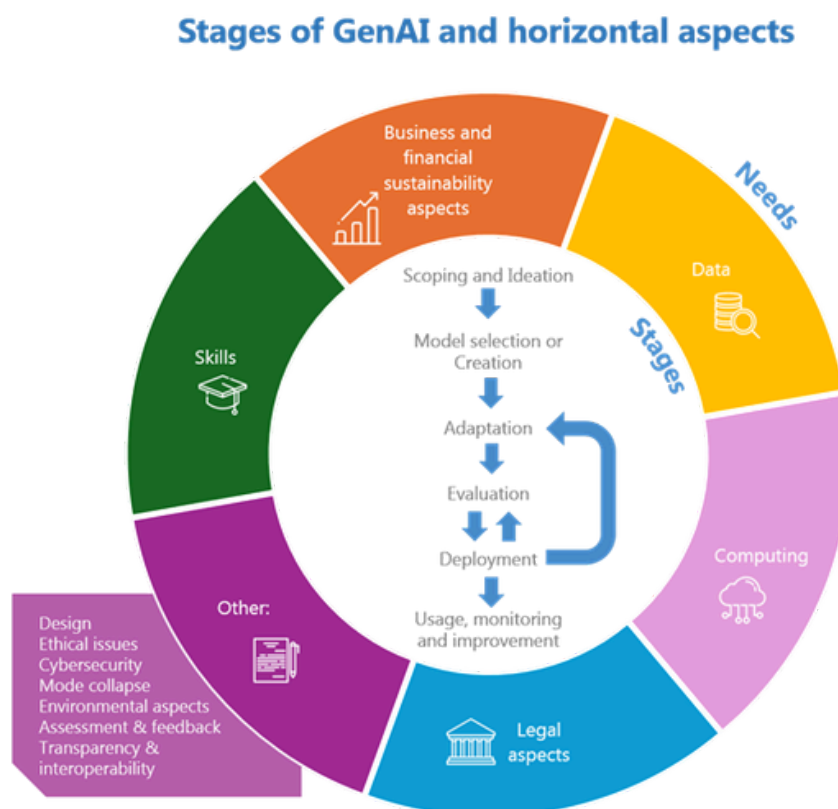


Figure 2 Challenges along the GenAI project lifecycle

The several stages can be described as follows:

- **Scoping and ideation:** The first phase of any GenAI project involves scoping, a phase where the objectives are defined based on business or scientific needs. During this phase, it's important to explore available resources against the requirements & desired outcomes. This includes discovering and investigating available datasets, existing AI models, popular model architectures, access to computing infrastructure, time, funding, legal requirements, needed skills, approaches used in similar scenarios (along with gaps and opportunities), identifying key actors involved, etc, Evaluation strategy, performance metrics and Key Performance Indicators (based on the scope, objectives and requirements) also need to be outlined in this phase although being measured during the evaluation phase.
- **Model selection or creation:** In this phase, the focus shifts to either selecting an existing AI model for the application or creating/pre-training a new one from scratch in order to better address the scoping phase requirements (note: the creation and deployment of a new model can be the original scope of a GenAI project, e.g., for model providers). Training a GenAI model from scratch is a very challenging and complex task as it requires a lot of data, computing, expertise, and resources (time and money). However, there are circumstances where the creation of a new model makes sense, e.g., to have full control over the model, tailoring it to specific needs from its conception for improved performance and operational reasons. This is an important implementation choice and part of an ongoing discussion over the merits of the different approaches.[21] Some recent reports also show a focus, in particular in industrial contexts, on the creation of smaller but capable models (built from scratch to perform in particular industrial contexts).[22]

[21]Ohm, Paul. "Focusing on Fine-Tuning: Understanding the Four Pathways for Shaping Generative AI." *Science and Technology Law Review* 25.2 (2024).

[22]https://assets.publishing.service.gov.uk/media/661e5a4c7469198185bd3d62/AI_Foundation_Models_technical_update_report.pdf

- **Adaptation:** This involves modifying an existing pre-trained model for specific tasks or to better perform in a specific context. Adaptation and/or fine-tuning (the distinction between these two terms seems somewhat blurry) use additional data to refine the pre-trained model's performance, particularly for specialised domains, use cases and tasks. Adaptation and fine-tuning can be conducted in several ways, depending on needs and focus. Different techniques require different data volumes and involve different training costs, technical difficulties and performance gains.[23] In recent literature we can observe Retrieval-Augmented Generation (RAG) is proposed as an adaptation technique as opposed to fine-tuning.[24] RAG works by fetching relevant documents or data snippets in response to queries, which are subsequently used to generate more accurate and contextually relevant outputs.[25] This technique (RAG) is integrated in application deployment and usage.
- **Evaluation:** In this stage the adapted model's performance is evaluated against quantitative and qualitative metrics and KPIs. This phase involves iterative testing to ensure that the model reaches its accuracy, relevance, reliability or scalability aims. The metrics used during the evaluation phase should be well aligned with the business/research objectives set (during the scoping phase) to showcase that the model is capable of generalising and avoiding biases while operating in real-world scenarios.
- **Deployment:** This phase involves integrating & deploying the model in an operational GenAI application. The attention shifts to performance monitoring and extraction of data-driven insights. For better reasoning (increasing relevance and optimising costs) RAG and prompt engineering methods can be employed. This stage involves the setting up of the infrastructure where the application will run, facilitating hosting, scaling, and managing of the application, ensuring its smooth operation and integration into existing systems.[26]

[23] Some techniques for adaptation and fine-tuning include (non-exhaustive list): supervised fine-tuning; zero or few-shots learning; parameter efficient fine-tuning (PEFT) such as LoRA; transfer learning, reinforcement learning from human feedback; instruction tuning, etc.

[24] e.g <https://www.acorn.io/resources/learning-center/rag-vs-fine-tuning> or <https://www.rungalileo.io/blog/optimi-llm-performance-rag-vs-finetune-vs-both>

[25] <https://www.acorn.io/resources/learning-center/retrieval-augmented-generation>

[26] <https://www.datascience-pm.com/the-genai-life-cycle/>

- **Usage, monitoring and improvement:** The use of the system or application, the continuous monitoring of the deployed application to track its performance, the measurement of user feedback, and the operational efficiency and improvements in models and systems are all important tasks in operational applications that require data exchange at different levels and between different stakeholders.

2.1 Data challenges

Data is a cornerstone of GenAI models, systems and applications and consequently throughout the various stages, distinct data needs and challenges emerge. The strategies chosen to effectively fulfil the application objectives affect the data needs, but at the same time, the needs, availability and cost of data will shape those strategies.

In most occasions, one or several stages for data investigation, acquisition, preparation and/or integration (e.g., real-time access to external data sources for RAG) are added. All these tasks are necessary, along the overall lifecycle and value chain of GenAI, although the needs, requirements and challenges are different at each stage.

- **Scoping and ideation:** During this stage comprehensive exploration and evaluation of available datasets (quantity, quality, diversity), governance aspects and costs is done. Requirements to access and exchange data is also an important factor to consider. Main challenges during this stage are data discovery, data access, unavailability of data (in quantity, quality or/and diversity) and restrictions to access existing private data. Regulatory challenges linked to data (Intellectual Property Rights, personal data, data governance) are very relevant and are tackled later in this chapter.
- **Model selection or creation:** If the option is to use a pre-trained existing model (the most common choice) benchmarking/evaluating existing models in smaller datasets and prompts typically facilitate the decision-making process. However, If the option of pre-training a new model is chosen, this entails dataset collection and preparation, cleaning, filtering, anonymising (if personal data is included), splitting of datasets for training/validation/testing etc., and model architecture design before conducting model training and validation. Dataset collection and preparation is quite a resource-consuming activity. **Very large, representative datasets are essential if the creation of a new model** is selected as the most appropriate option. These datasets will be used for both training and validation and must encapsulate the characteristics that the generative model intends to learn.

Data quality and diversity within the datasets are important factors to ensure that the model is robust and can learn patterns, reduce biases and generalise. The main data challenge at this stage is quantifying, sufficient diversity and resources for preparation of the data.

- **Adaptation:** Smaller but more relevant task/case-specific curated datasets are typically needed to fine-tune/validate the model and adjust its hyperparameters optimally as needed. This ensures that the model can be used in specialised tasks achieving optimal performance. The main data challenge in this stage is quality (in all its dimensions). [27]
- **Evaluation:** Diverse, task-specific evaluation datasets (with datasets not used during model training or fine-tuning) representing data that better reflect the deployment environment are required to evaluate the model's performance before its operational use. Prompt datasets could also be needed during the evaluation phase for testing the model's ability to respond effectively to different types of inputs. Depending on the evaluation aims and strategy, different datasets might be needed (e.g. for bias detection, fairness or security-related aspects). The main challenge at this stage is data quality and representativeness.
- **Deployment:** If RAG methods are used, external knowledge sources in the form of unstructured long-form texts, structured database records or semi-structured files are needed to refine the model's response in relevance terms. The main data challenge in this stage is the real-time access of relevant data. Other challenges include standard Application Programming Interfaces (APIs) and data models.
- **Usage, monitoring and improvement:** Collected feedback and data stemming from user interactions could be exploited and analysed for the model's iterative maintenance and evolution. This will ensure that the model and consequently the GenAI application remains effective, relevant and aligned with the business goals over time. The main data challenges relate to data privacy (in the interactions between the user and the application, or the feedback provided) and access to real-life and real-time data.

[27] BDVA (2024), Elevating Data Quality A Paradigm Shift for Data Spaces and AI Needs, May 2024, version 1.0; CC BY-SA 4.0, available at <https://bdva.eu/news/new-paper-%F0%9F%91%89-elevating-data-quality-a-paradigm-shift-for-data-spaces-and-ai-needs>

2.2 Computing needs

GenAI models have redefined what is considered feasible in machine learning. Billions of model parameters are tuned using trillions of data tokens. This tedious training requires thousands of Graphics Processing Units (GPUs) processors and months of running time.

The needs of GenAI have led to massive investments in GPU-powered AI supercomputers. GPU processors are the central element, as their capacity in tensor calculation required by deep neural networks is superior to more traditional Central Processing Units (CPUs). However, due to the extreme size of GenAI models, single GPU processors are not sufficient for training them. High-speed interconnection networks are required to effectively employ hundreds or thousands of GPUs simultaneously. Obviously, data-intensive AI workloads also require high-performance data storage solutions. However, typically storage systems are not the bottleneck in the context of supercomputing systems.

Training large AI models from scratch, also called pre-training, is not the only type of GenAI workload with non-trivial computing needs. Further adaptation of the models via fine-tuning places similar demands on the computing infrastructure, albeit on a smaller scale. Inference, the final end use of the AI models, on the other hand poses different needs. When large models, such as LLMs, are provided to users often a decisive factor is the size of the models. Largest models can require several GPUs just to gather enough memory to hold a single instance. As inference services serve a large number of users the memory requirements can be challenging. Various optimisation techniques and more space-efficient model architectures can be used to reduce memory requirements significantly, with some sacrifice in quality of the outputs.

Training of very large AI models is demanding, but on the other hand it is a fairly straightforward bulk operation from an architectural point of view. However, as AI models are being integrated more deeply into diverse IT systems, the architectural demands are also amplified. In a modern environment the components of an AI system can be transparently deployed over cloud, High Performance Computing (HPC) and edge computing platforms. It can be envisioned that AI systems will be an integral part of the computing continuum of the future.

2.3 Skills

AI transformation has created needs for a wide range of new skills across the workforce. Here we focus strictly on the more specific data-related skills needed for Gen AI. These skills needs can be mapped against the different stages of any GenAI life cycle.

- **Scoping and ideation:** Data management skills are essential when setting the scope. Skills range from expertise in technical data storage systems to data curation and data science skills required for understanding the statistical aspects such as biases inherent in the data. Integration of data sources requires an understanding of the content of data as well as the technical skills required for access. Additionally, legal competence is required for compliance, IPR and data protection topics, inter alia.
- **Model selection or creation:** Pre-training of AI models is a new type of challenge, giving rise to a completely new job profile of a machine learning engineer. Designing model architecture, designing training processes, monitoring, optimising and scaling of training computing runs are all part of this new skillset. If an existing model is selected, then a similar skillset is required, but the requirements for training process related skills are less significant when using existing trained models.
- **Adaptation:** The skillset of the pre-training phase remains relevant, but much more diverse skills are needed at the adaptation stage, especially for model alignment. Legal, ethical, socioeconomic and other skills are necessary to identify and formulate points of views that should be addressed in model alignment. The skill requirements of this phase are perhaps the ones that require further assessment.
- **Evaluation:** Model evaluation builds on the goals set in selection/pre-training and adaptation phases, so the same skill needs apply. In addition, the evaluation of models with human feedback, or of a deployed model would benefit from expertise in human-computer interaction and user interface design.

- **Deployment:** In the final phase the model is deployed for use, also called inference. Especially if the user base is large then advanced data engineering, cloud architecture and distributed software development skills are required. Additionally, machine learning skills are needed to put in place monitoring to follow the performance of the deployed models.
- **Usage, monitoring and improvement:** The users of GenAI models require practical AI skills, most importantly the capability to write prompts (prompt engineering). Also, at least a basic understanding of AI methods and their limitations is needed. Data privacy and data protection become relevant skills for this stage as well.

In all stages strong legal-tech expertise is needed.

The skill needs of GenAI are quickly evolving as the technology develops and matures. It is foreseeable that solution providers will grow their offering to cover many aspects of the model lifecycle, meaning that GenAI users can outsource skills more easily. This development obviously creates the additional skill needed for planning and implementing procurement successfully.

2.4 Legal aspects

Recent advancements in AI such as general-purpose AI (GPAI) and Gen necessitate careful consideration of the many legal dimensions that are triggered by this set of complex technologies. Given that GPAI/GenAI models and systems are typically trained on extensive datasets, several legal challenges arise, including issues related to data governance, privacy and data protection, intellectual property rights (IPR), cybersecurity, competition, and fundamental rights. This section examines a selection of these key legal topics that are essential for fostering high-quality and trustworthy AI, addressing in particular data governance regulation, regulation of IPR, privacy and data protection regulation and cybersecurity regulation. The analysis focuses primarily on adherence to the European Union's Artificial Intelligence Act (AI Act), which forms a crucial part of the broader regulatory landscape.[28] While ethical considerations are inherently connected to these legal aspects, they will be discussed in a separate section (see section 2.6 addressing 'Some other relevant challenges') - this is because law is a formal, codified system of rules and regulations established by a governing authority and benefiting from enforcement mechanisms, whereas ethics are generally understood as principles and guidelines based on societal and personal values, and can be subjective and susceptible to changes.

[28] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act).

2.4.1 The EU AI Act - a comprehensive approach to regulating AI

The AI Act aims to address the unique challenges of AI systems and GPAI models, including GenAI, with a set of specific legal requirements. It is worth noting that this field is dynamic and rapidly evolving, so the below legal analysis does not purport to be fully exhaustive, but rather illustrative of a disciplinary and methodological direction, and new issues may arise over time. The way in which the AI Act (and the forthcoming General-Purpose AI Code of Practice[29]), is enforced and interpreted over time will be pivotal. The complexity, opacity, lack of causality, unpredictability and degree of autonomy displayed by AI systems and GPAI models also introduce challenges in predictability and legal compliance.[30], [31]

Definition of GenAI in the AI Act:

In the AI Act the concept of GenAI is captured in the definition of an AI system, as it refers to the generation of ‘outputs such as content’.[32] For deployers of AI systems ‘that generate or manipulate image, audio or video content constituting a deep fake’, there are specific transparency obligations.[33] However, most of the obligations related to GenAI are expected to fall within the scope of GPAI models and systems, including GPAI models with systemic risks.

[29]<https://digital-strategy.ec.europa.eu/en/news/ai-act-participate-drawing-first-general-purpose-ai-code-practice#:~:text=The%20Code%20of%20Practice%20will,of%20Practice%20to%20demonstrate%20compliance>

[30] Novelli, C., Casolari, F., Hacker, P., Spedicato, G., & Floridi, L. (2024). Generative AI in EU law: liability, privacy, intellectual property, and cybersecurity. arXiv preprint arXiv:2401.07348.

[31] Fernández Llorca, D., Charisi, V., Hamon, R., Sánchez, I., Gómez, E., ‘Liability Regimes in the Age of AI: a Use-Case Driven Analysis of the Burden of Proof’ (2023), Journal of Artificial Intelligence Research, Vol. 76, 12 March 2023.

[32] Art 3(1) AI Act.

[33] Art 50(4) and recital 134 AI Act.

A GPAI model is defined as ‘an AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications’.[34] This definition is applicable to GenAI models,[35] which are considered as a type of GPAI models, ‘given that they allow for flexible generation of content, such as in the form of text, audio, images or video, that can readily accommodate a wide range of distinctive tasks’.[36] The AI Act introduces rules for GPAI models and systems, which will apply from August 2025. [37] The Act aims to ensure transparency along the value chain and address possible systemic risks of the most capable models.

Approach to regulating GPAI/GenAI in AI Act:

The AI Act regulates GenAI in a variety of ways and in different forms. We highlight the following regulatory categories as key concepts in the definition of the legal framework and briefly explain how they relate to specifically to GenAI models:

- *General-purpose AI (GPAI) models* – specific provisions are introduced for GPAI models, mostly aimed at increasing transparency of these models,[38] including the training datasets; as well as compliance with copyright and ensuring information sharing along the AI value-chain. It is important to understand that GPAI models are generally assumed to be integrated into AI systems and have downstream applications. As mentioned above, within the AI Act, GenAI models are described as a subtype of GPAI models,[39] and the obligations to GPAI models can therefore be considered to apply.

[34] AI Act Art 3(63).

[35] For the legal definition of ‘General-Purpose AI model’ see AI Act Art 3(63). Also see discussion in Fernández-Llorca, D., Gómez, E., Sánchez, I., & Mazzini, G. (2024). An interdisciplinary account of the terminological choices by EU policymakers ahead of the final agreement on the AI Act: AI system, general purpose AI system, foundation model, and generative AI. *Artificial Intelligence and Law*, 1-14.

[36] AI Act recital 99.

[37] For more details, see Art 113 AI Act.

[38] AI Act, Article 53 and Annex XI.

[39] AI Act recital 99.

- *General-purpose AI models with systemic risks* – When GPAI models are considered to pose systemic risks, additional obligations apply, that are aimed at identifying and mitigating these risks. The notion of systemic risks is linked with the capabilities of the model and their potential significant impact on the Union market in terms of reach or ability or negative effect on public health, safety, public security, fundamental rights, or the society as a whole. GenAI models could potentially fall under this category, and a presumption is raised based on the cumulative amount of computation for their training.[40] This regulatory category is expected to apply to only a small subset of the most advanced GPAI models, including the most advanced GenAI models.
- *High-Risk AI systems* – The AI Act introduces stricter measures for high-risk AI systems, including in the areas of data governance, record keeping, rigorous adversarial testing, technical documentation, compliance management, explaining of AI decisions, informed consent, and monitoring. The high-risk nature of a system depends on the downstream use that is made of it (e.g., when it is part of a product that requires a conformity assessment).[41] GPAI (including GenAI) models could be considered as AI systems if additional components are added – e.g., a user interface - or if they are integrated into, and form part of another AI system.[42] For this reason, these rules may be directly applicable to GenAI/GPAI systems - GPAI systems may be used as high-risk AI systems by themselves or be components of other high-risk AI systems.[43]

[40] Art 51(2).

[41] For classification rules, see AI Act Art 6, complemented by Annex III.

[42] AI Act recital 97. For the legal definition of a GPAI system, see Article 3(66).

[43] AI Act recital 85.

- *Certain AI systems* – special transparency obligations are imposed on ‘certain AI systems’, a term that includes for example ‘AI systems, including general-purpose AI systems, generating synthetic audio, image, video or text content’ or ‘an AI system that generates or manipulates image, audio or video content constituting a deep fake’. Providers of these AI systems are required to ‘ensure that the outputs of the AI system are marked in a machine-readable format and detectable as artificially generated or manipulated’.[44] In addition, deployers of AI systems that generate content constituting deep fakes are required to disclose that the content has been artificially generated or manipulated.[45] The same obligations apply to deployers of an AI system that ‘generates or manipulates text which is published with the purpose of informing the public on matters of public interest’. These obligations have particular relevance for GenAI.

The AI Act introduces, inter alia, rules for the placing on the market, the putting into service, and the use of GPAI systems and models.[46] It regulates the conduct of providers and deployers of these systems. It is important to be familiar especially with the notion of provider, given its relevance for GPAI systems and models. The AI Act defines a ‘provider’ as a natural or legal person, public authority, agency or other body that develops a general-purpose AI model or that has a general-purpose AI model developed and places it on the market or puts the AI system into service under its own name or trademark, whether for payment or free of charge’.[47] In some circumstances (e.g., when a high-risk AI system is a safety component of a product), the product manufacturer will be considered a provider of the high-risk system.[48]

[44] AI Act Art 50(2). Also see Hamon, R., Sanchez, I., Fernandez Llorca, D. and Gomez, E., Generative AI Transparency: Identification of Machine-Generated content, European Commission, Ipsra, 2024.

[45] AI Act Art 50(4).

[46] See AI Act Art 1(2)(a), (c) and (e).

[47] AI Act Art 3(3).

[48] AI Act Art 25(3).

2.4.2 Data Governance regulation

Data governance is a key component of AI (and GPAI/GenAI) governance and can be defined as the set of rules, policies, relations, decision-making structures and processes established among different kinds of actors to collect, share and use data. [49] In broad terms, it refers to the exercise of authority and control over the management of data. Provisions on data governance for GPAI/GenAI include related aspects such as data protection and privacy, which are dealt with separately in 2.2.4. While several other horizontal EU regulations in this area are also relevant (e.g. the Data Governance Act and the Data Act - see Annex A), as well as some sectoral regulations, in this section, we concentrate in particular on the conceptualisation of data governance as set forth in the AI Act. In accordance with Recital 27 of the Act, data governance and management refer to the "processing of data that meets high standards in terms of quality and integrity".

Since GPAI/GenAI models are typically trained on large amounts of data, the correct governance of this data is important. Data governance is a continuous process that spans the entire GenAI lifecycle, but it is particularly critical during the phases of data collection and preparation, training, and validation/testing. These stages lay the foundation for the performance of a GPAI/GenAI system and for its compliance with regulatory requirements. Ensuring high standards of data quality and integrity at these points is therefore essential for the development of trustworthy, reliable, and legally compliant GPAI/GenAI models and systems.

[49] See Farrell, E., Minghini, M., Kotsev, A., Soler Garrido, J., Tapsall, B., Micheli, M., Posada Sanchez, M., Signorelli, S., Tartaro, A., Bernal Cereceda, J., Vespe, M., Di Leo, M., Carballa Smichowski, B., Smith, R., Schade, S., Pogorzelska, K., Gabrielli, L. and De Marchi, D. (2023), European Data Spaces - Scientific Insights into Data Sharing and Utilisation at Scale, Publications Office of the European Union, Luxembourg; also see <https://www.oecd.org/en/topics/sub-issues/data-governance.html>

- **Relevant provisions in the AI Act**

The AI Act explicitly links data governance to AI governance, highlighting it as one of the major requirements of a trustworthy AI system, including the ability to trace how data is used throughout the lifecycle. It governs a wide range of data categories in its regulation of AI systems, covering both personal and non-personal data, and specifically including content data used in the training of these systems. In particular, the governance and management of high-risk AI systems requires stringent adherence to data quality and governance standards, as highlighted in Article 10 (Data and data governance), Article 11 (Technical documentation), and Article 12 (Record-keeping).

GenAI/GPAI models are often sourced from a variety of datasets with little available documentation on the training data or data collection practices, which may arguably fail to properly address issues including consent, authorisation, compensation, attribution or regard for other types of legal rights or contractual obligations.[50] Furthermore, many of the entities developing GPAI/GenAI models and systems provide sparse information about how they constructed, filtered and organised their underlying data sets, as this is often valuable commercial information.[51] It is therefore important to be able to establish data provenance as a means of ascertaining if the datasets underpinning GPAI/GenAI models and systems are accurate, complete and representative. In this regard, and as previously mentioned, the AI Act emphasises the need for transparency and documentation, entailing the provision of information to support the understanding of AI systems and data throughout their lifecycle.[52]

[50] Longpre, S., Mahari, R., Obeng-Marnu, N., Brannon, W., South, T., Kabbara, J., & Pentland, S. (2024). Data Authenticity, Consent, and Provenance for AI Are All Broken: What Will It Take to Fix Them?. <https://doi.org/10.21428/e4baedd9.a650f77d>

[51] Aaronson, S. A. (2024). Data Disquiet: Concerns about the Governance of Data for Generative AI. Available at SSRN 4784412.

[52] See Hupont, I., Micheli, M., Delipetrev, B., Gómez, E., & Garrido, J. S. (2023). Documenting high-risk AI: a European regulatory perspective. *Computer*, 56(5), 18-27. and 25, 56 (2023), available at: <https://link.springer.com/article/10.1007/s10676-023-09725-7>

The reliance of many Large Language Models (LLMs) on large-scale data sets can also lead to data quality and validity concerns, aspects that play a key role in the interoperability, reliability, trustworthiness and usability of data. As mentioned, the AI Act includes data quality and accuracy criteria in Article 10, specifying that “High-risk AI systems which make use of techniques involving the training of AI models with data shall be developed on the basis of training, validation and testing data sets that meet the quality criteria referred to in paragraphs 2 to 5 whenever such data sets are used.” Recital 44 also notes that “datasets [...] should be relevant, sufficiently representative, and to the best extent possible free of errors and complete in view of the intended purpose of the system.”

Compliance challenges: These important AI and data governance rules may increase the legal compliance costs of GPAI/GenAI providers, who may face challenges in maintaining up-to-date technical documentation and compliance records due to the complexity of data sourcing. The lack of a structured environment for data sharing and management further complicates the ability of GenAI providers to ensure the integrity, provenance, and transparency of the data used in model development.

2.4.3 Regulation of Intellectual Property Rights (IPR)

The data on which GPAI models are trained may be protected by copyright (when for instance the training material qualifies as copyright works) as well as by other forms of exclusivity. AI certainly presents enormous innovation opportunities; however, it also poses challenges to the creative economy, especially authors and creators. The training of GPAI models is an activity that needs to comply with the applicable legal framework, which, depending on the specific case, may require forms of authorisation, either statutory (exceptions or limitations) or contractual (licences, agreements, etc).

The relevance of copyright law for the training of GPAI/GenAI models is justified by two considerations. First, the level of creativity needed for a work to attract copyright protection is not particularly demanding: - they are protectable as long as they are the author's own intellectual creation thus implying a degree of free and creative choice.[53] This definition likely covers most texts, photographs, and other audiovisual or software content available online and used for training these systems - it may also include original databases. Specific rules apply to databases protected by the sui generis database right (SGDR), which are not discussed here.

A second element is the scope of the rights conferred by copyright. Particularly relevant here is the reproduction right, defined in EU law as covering "direct or indirect, temporary or permanent reproduction by any means and in any form, in whole or in part".[54] As a result of this broad definition, many of the copies made during training activities may trigger an infringement of the right of reproduction.

[53] Case C-5/08 Infopaq para 39.

[54] Info Soc Art 2.

Compliance challenges: The regulatory burden on AI developers in carrying out due diligence on the training data is certainly demanding. Identifying the rightsholders for each individual work is sometimes difficult, particularly for online User-generated content. Moreover, the question of whether a work is protected by copyright or whether a specific activity is covered or not by an exception is sometimes difficult to answer without an individual assessment and access to legal advice, especially when the level of expressive freedom necessary to claim copyright is limited (e.g., metadata describing an image). When considering the billions of data used for training, or even the more limited datasets used for fine-tuning, the cost of compliance may be high.

Articles 3 and 4 of the **Copyright in the Digital Single Market Directive (CDSM)**[55] introduced a regulatory regime whereby reproductions and extractions in the context of text and data mining do not require the permission of rightsholders, as long as the conditions set out in the respective provisions are satisfied. Without entering into details,[56] it can be broadly stated that Article 3 CDSM benefits research organisations and cultural institutions that train GPAI models for the purpose of scientific research, whereas Article 4 also covers commercial uses irrespective of the nature of the party carrying out the training. We focus on the conditions for the application of Article 4, as this is likely to be the most frequent scenario in the context of commercial AI training.

The main conditions that need to be complied with are:

- a) Under Article 4(3), the rightsholder retains the option to reserve the rights and thus make the exception inapplicable - in other words, to opt out of the exception. Such reservation needs to be made by expressing it in an explicit and machine-readable manner.
- b) The content must be lawfully accessible – while a lack of clarity persists concerning the interpretation of ‘lawfully accessible’, a non-exhaustive list of possible legal bases of lawful access may include licensed content (whether open access or as part of subscriptions), as well as content freely available online.[57]

[55] Copyright in the Digital Single Market (2019).

[56] See Margoni, T., & Kretschmer, M. (2022). A deeper look into the EU text and data mining exceptions: harmonisation, data ownership, and the future of technology. *GRUR international*, 71(8), 685-701.

[57] See CDSM recital 14.

What is described above can be considered as a general overview of the main elements to take into account when looking at copyright compliance. The AI Act introduces measures that complement the above copyright acquis without nonetheless altering it. These rules are binding on providers of GPAI models and are therefore part of the compliance with the AI Act.

Two obligations of the AI Act are especially important in this regard:

- **Setting up a policy to comply with Union law on copyright** - Under Article 53(1)(c), Providers of GPAI models need to put in place a policy to comply with EU Union law on copyright and related rights, in particular, to identify and comply with the opt-out of rightsholders under Article 4(3) of the CDSM. This policy will have to be documented and may also need to be shared with the AI Office under Article 53(1)(a) to assess compliance with the AI Act.[58]
- **Provide a detailed summary on the pre-training dataset** - Article 53(1)(d) requires providers of GPAI models to draw up and make publicly available a sufficiently detailed summary of the content used for training the GPAI model. The text further specifies that the AI Office will have to provide a template for this. The purpose is to allow parties with legitimate interests "to exercise and enforce their rights".[59]

[58] See also Annex XI and Annex XII respectively. As a minimum information, the licence for the model as well as the training process will have to be made available, and the copyright policy will likely have to be covered at least indirectly.

[59] AI Act recital 107.

2.4.4 Privacy and Data Protection regulation

The vast amount of data (text, images, videos etc.) used to train GenAI may contain personal data. As outlined by the European Data Protection Supervisor (EDPS), personal data processing in a GenAI system can take place at various levels and stages, without necessarily being immediately obvious, including during the creation of training datasets, during the training stage itself, by inferring new or additional information once the model is created and in use, or through the inputs and outputs of the system once it is in operation.[60]

Under the General Data Protection Regulation (GDPR), personal data is defined as information relating to an “identified or identifiable” individual. An “identifiable” individual is considered to be an individual who can be identified either directly or indirectly, in particular by reference to identifiers such as a name, an identification number, location data, an online identifier, or factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of the individual in question (Art. 4). The processing of such data, i.e. personal data, should always adhere to the data protection principles proclaimed by the regulation (Article 5). In other words, personal data must be:

1. processed lawfully, fairly and in a transparent manner in relation to the individuals concerned (lawfulness, fairness and transparency).
2. collected for specified, explicit and legitimate purposes, and not be further processed in a manner that is incompatible with the initial purpose (purpose limitation).
3. processed only when necessary for the stated purposes (data minimisation).
4. kept accurate and up to date (accuracy).
5. stored only for as long as necessary for the specified purposes of processing (storage limitation).
6. processed in a manner that ensures appropriate data confidentiality and security (security and confidentiality).

[60] The European Data Protection Supervisor has provided guidance for ensuring data protection compliance when using GenAI systems - see EDPS (2024), ‘Generative AI and the EUDPR. First EDPS Orientations for ensuring data protection compliance when using Generative AI systems.’

These principles should also be reflected in data processing for the purposes of developing and deploying GenAI systems. As for the lawfulness principle, any data processing requires a legal basis, i.e., a specific grounds for processing (listed in Article 6). Grounds of processing relevant to the training and use of GenAI systems include consent given by the concerned individual, the necessity of data processing for the performance of a contract, and the necessity of data processing for the purpose of the legitimate interest of the data controller or a third party. The lawfulness of each step of data processing in GenAI development and use must be assessed individually.[61]

When it comes to the principles of purpose limitation and data minimisation, it is necessary to have a reflection on the purposes to be pursued by the data processing as well as on the necessity of data being processed for those purposes, when designing a GenAI system.[62]

Once in use, personal data used for the training of GenAI systems should be processed in a secure manner (principle of confidentiality and security) and must only be stored for as long as necessary for the earlier specified purposes of processing (principle of storage limitation).

According to the principle of accuracy, personal data must be kept accurate and reasonable efforts should be used to ensure this. To address the possibility of GenAI models “inventing” facts or producing so-called hallucinations, developers may inform users of the possibility that some output data may be factually incorrect, along with providing opportunities to suggest corrections.[63]

The GDPR also includes the principles of data protection by design and by default (Article 27), which aim to protect personal data throughout the entire lifecycle of data processing. By complying with this principle and adopting a risk-oriented approach, the risks that GenAI may entail can be considered and mitigated in advance.

[61] See e.g. Bartels, M. (2024). A Balancing Act: Data Protection Compliance of Artificial Intelligence. *GRUR International*, 73(6), 526-537.

[62] For example, the French Data Protection Authority (CNIL) has provided some guidelines on what dimensions to consider when identifying the personal data strictly necessary (see Taking into account data protection when designing the system, CNIL, 2024).

[63] Bartels (2024), p. 536.

Along with the GDPR and its principles, the rights to privacy and data protection are also addressed in the AI Act. Indeed, Article 2(7) of the AI Act confirms that the GDPR applies to personal data processed in connection to AI development and use. Additionally, as stated in recital 69 of the Act: “The right to privacy and to protection of personal data must be guaranteed throughout the entire lifecycle of the AI system. In this regard, the principles of data minimisation and data protection by design and by default, as set out in Union data protection law, are applicable when personal data are processed.” Further, the Act enshrines “privacy and data governance” as one of the principles applicable to all AI systems (Recital 27).

Compliance challenges: as pointed out by the European Data Protection Supervisor, some legal questions remain open relating to the processing of personal data and compliance with data protection rules when developing and using GenAI systems. Additional questions may arise as use of these systems increases and technology evolves. Where a GenAI developer or provider claims their model or system does not process personal data, it is important to understand what steps or procedures were taken to ensure this is not the case. Data controllers are responsible for and accountable for implementing appropriate technical, organisational and procedural measures to ensure the effective exercise of the rights mentioned above. These measures should be designed and implemented from the early stages, allowing for detailed recording and traceability of processing activities.

2.4.5 Cybersecurity regulation

The AI Act currently only mandates cybersecurity measures for high-risk AI systems (Article 15) and for GPAI models with systemic risk (Art. 55). According to Article 55 (c) of the Act, providers of GPAI models posing systemic risks (i.e., with high-impact capabilities) must fulfil standard obligations and additionally conduct model evaluations, including adversarial testing (red teaming); assess and mitigate risks; document and report incidents to the AI Office; and maintain adequate cybersecurity protection. In order to ensure compliance, a security risk assessment should be conducted taking into account the design of the system, to identify risks, and implement the necessary mitigation measures. This process requires an integrated and continuous approach using proven cybersecurity practices and procedures, combined with AI-specific controls. The cybersecurity requirement is further explained in Recital 76 of the AI Act.

Provisions in the EU Cyber Resilience Act (CRA) will also apply in varying degrees to GenAI (e.g., Articles 3(1), 6 and 7 affecting Products with Digital Elements). The NIS2 Directive expands cybersecurity measures to cover additional sectors and entities, making it relevant for those that make use of GenAI, including the digital infrastructure and services sectors.

Compliance challenges: The use of GenAI systems may amplify existing cybersecurity risks or create new ones. As risks evolve quickly, regular monitoring and updates of the risk assessment are needed, as well as specialised technical support.

2.5 Business and financial sustainability aspects

Implementing GenAI brings up important business and financial considerations, as it requires significant financial investments and involves multiple actors across its value chain. Key costs include access to infrastructure, research investments, data acquisition and human resources for development activities (considering technical, software, business, regulatory compliance aspects, etc.) need to be acknowledged.

To ensure their long-term business and financial sustainability, actors involved in these developments need to consider funding needs and establish viable monetisation strategies. Depending on the project's scope (e.g., research, commercial), various funding and investment mechanisms can be employed at early stages of development. This include public or private funding used for the creation of new models, adaptations, conducting research etc. When it comes to monetisation strategies important dimensions need to be considered including the openness of the GenAI model (Open public AI models that encourage usage and contributions by the community versus closed proprietary models with restricted access, that are used for offering competitive advantage) and how coupled the GenAI offering is to the business model (e.g., direct monetisation of AI offering through subscriptions, licences, usage-based or other model type vs Indirect monetisation through integration of the AI offering into existing bundle of products or services).

2.6 Some other relevant challenges

As indicated in the introduction to this section, many other challenges apply to GenAI. While it is not the purpose of this paper to enter in depth in all of them, we address some that are connected (or potentially linked) to data aspects.

- **Design challenge:[64]** Although the prevalent generative modelling paradigms such as diffusion models[65] and LLMs[66] demonstrate exceptional sample quality, there are multiple significant problems related to the generative process design. For example, the multi-stage denoising procedure in diffusion models is iterative, which slows down inference significantly and frequently necessitates hundreds to thousands of network function evaluations (NFE) in order to produce high-quality samples[67],[65]. Similar to this, the sequential token generation caused by the autoregressive structure in LLMs can result in sluggish inference. This is not the case for other generative models, such as variational autoencoders (VAEs) and Generative adversarial network (GANs), which only need one NFE yet have issues with mode collapse[66] and blurry sample generation.[68]

[64] Manduchi, L., Pandey, K., Bamler, R., Cotterell, R., Däubener, S., Fellenz, S., ... & Fortuin, V. (2024). On the challenges and opportunities in generative ai. arXiv preprint arXiv:2403.00025. <https://arxiv.org/abs/2403.00025>

[65] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. Advances in neural information processing systems, 33, 6840-6851.

[66] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. & others (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

[67] Arjovsky, M., Chintala, S., & Bottou, L. (2017, July). Wasserstein generative adversarial networks. In International conference on machine learning (pp. 214-223). PMLR.<https://proceedings.mlr.press/v70/arjovsky17a.html>.

[68] Dosovitskiy, A., & Brox, T. (2016). Generating images with perceptual similarity metrics based on deep networks. Advances in neural information processing systems, 29.

- **Ethical issues:[69]** The usage of GenAI technology gives rise to ethical questions, especially when those uses could be malevolent, manipulative or likely to cause harm. Deepfake technology, for example, can produce extremely life-like, but fake, content, which could spread false information or cause harm. It is necessary to design policies, rules, and safeguards to ensure the ethical and responsible usage of GenAI models. This entails putting strategies in place to identify and lessen the misuse of GenAI, to encourage accountability and transparency, and to deal with any potential biases in the outputs that are produced. Other ethical issues may arise in relation to the data inputs and processing, as well as the model design, e.g., bias, representativeness, accuracy, fairness, propagation of stereotypes, discrimination, safety, etc. Some of these issues may be addressed in legislation but there is also a need for robust testing, ongoing monitoring and participatory processes to help mitigate these risks. Explainability is also an important aspect.
- **Mode collapse:[70]** This is the result of a generative model producing limited or repeating variations rather than fully capturing the diversity of the training set. [71] An image creation model, for instance, might reliably generate pictures of a particular object while excluding other potential objects included in the training set. One important difficulty in the field of GenAI research is overcoming mode collapse. To help the model provide a wider range of outputs, strategies like enhancing model topologies, optimising loss functions, or utilising ensemble approaches are investigated.

[69] Dwivedi, Y.K.; Kshetri, N.; Hughes, L.; Slade, E.L.; Jeyaraj, A.; Kar, A.K.; Baabdullah, A.M.; Koohang, A.; Raghavan, V.; Ahuja, M.; et al. "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *Int. J. Inf. Manag.* 2023, 71, 102642.

[70] Zhang, C., Zhang, C., Zhang, M., & Kweon, I. S. (2023). Text-to-image diffusion models in generative ai: A survey. *arXiv preprint arXiv:2303.07909*.

[71] For further reading on this, see Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, 631(8022), 755-759.

- **Content pollution and quality degradation:** There is a growing concern about the potential for AI-generated content to pollute future training datasets, leading to a vicious cycle where AI models are increasingly trained on content produced by other AI models rather than human-created content.[72] This could result in a gradual degradation of content quality and nuance over time, as subtle human insights and creativity may be lost or diluted through repeated AI processing and regeneration.[73] Ensuring diverse, high-quality human-generated content remains a core part of training datasets will be crucial to mitigate this risk.[74]
- **Transparency and interpretability:[75]** A lot of GenAI models, especially deep neural networks, are sometimes referred to as "black boxes," which means that it is difficult for humans to understand how they make decisions. Trust can be hampered by this lack of interpretability and transparency, particularly in important areas where justifications and explanations are necessary. In order to shed light on the inner workings of generative models, researchers are actively investigating ways to improve their interpretability. Some of these strategies include attention processes, visualisation techniques, and producing explanations in addition to the model's outputs.
- **Assessment and feedback:[75]** Assessing the level of originality and quality of generative products is a challenging issue. Semantic relevance, coherence, and novelty are examples of desirable features of generated information that may not be fully captured by traditional evaluation criteria. Research on creating trustworthy evaluation criteria that are particular to GenAI is still continuing. It is also essential to get insightful input from experts or consumers in order to iteratively enhance the models. While difficult, gathering large amounts of feedback and assimilating it into the training process is essential for improving the model.

[72] Rob Kniaz, The Incoming Tidal Wave Of Data Pollution In AI, Forbes, May 9, 2023. <https://www.forbes.com/sites/robkniaz/2023/05/09/the-incoming-tidal-wave-of-data-pollution-in-ai/>

[73] AI produces gibberish when trained on too much AI-generated data, Nature, September 2024. <https://www.nature.com/articles/d41586-024-02355-z>

[74] Gibney, E. (2024). AI models fed AI-generated data quickly spew nonsense. Nature, 632(8023), 18-19.

[75] Evans, O., Cotton-Barratt, O., Finnveden, L., Bales, A., Balwit, A., Wills, P., ... & Saunders, W. (2021). Truthful AI: Developing and governing AI that does not lie. arXiv preprint arXiv:2110.06674.

- **Cybersecurity measures to consider in different parts of the lifecycle:** Cybersecurity companies and researchers have emphasised a holistic approach to AI trustworthiness. This includes at least two key aspects that relate to the lifecycle of AI as well as AI trustworthiness definitions and attributes. The former focuses on all the aspects of the AI lifecycle. For instance, in the phase of fine-tuning or (re-) training, data poisoning is a threat to AI trustworthiness. A data poisoning attack “aims to modify a training set such that the model trained using this dataset will make incorrect predictions.”[76]
- **Environmental aspects:** The use of GenAI while revolutionary in many aspects, poses important environmental challenges that should be acknowledged. One major concern is associated with the computational power required to train and run large GenAI models. These processes consume significant resources, demanding not only substantial energy but also the exploitation of other natural resources, such as water for cooling data centres and materials for the production of specialised hardware (i.e. GPUs) which also account for increased e-waste. On top of this, intensive or unpredictable usage (leading to shifts of patterns in resources demands) and the iterative way that many GenAI tools are operating — producing multiple images or scenarios in cycles—further amplify their environmental impact. Each iteration adds to the overall energy consumption, leading to an increase in the carbon footprint. Given that much of the energy powering these systems comes from non-renewable sources, this growing demand raises environmental sustainability concerns and the necessity for mitigation measures (e.g., working with smaller models).

[76] source: <https://blog.f-secure.com/data-poisoning-in-action/>

3 The transformative potential of data spaces for Generative AI

Data spaces can offer substantial value to GenAI by addressing recognised challenges and unlocking new opportunities. This chapter examines this potential, presenting data spaces as a ‘solution’ to some of the issues, challenges, and needs outlined in chapter 2, while also highlighting emerging opportunities. Part 2 of this whitepaper offers a collection of learnings from use cases and examples gained from the usage of Gen models and applications in several sectors, which is also very relevant in this context.

Data spaces can support the development and utilisation of trustworthy GenAI through a set of tangible benefits to the stakeholders involved. Central to this is their **ability to provide the essential foundations for secure and efficient data sharing**, enabling GenAI to access diverse, domain-specific (including industrial) datasets. Next to this, the **digital ecosystem perspective** that data spaces bring could be seen as an attribute of equal importance. Data spaces cultivate an ecosystem of innovation and collaboration allowing different stakeholders to be part of the value chain and unfold new business opportunities. Moreover, Data Spaces serve as important **enablers for compliance with legal and regulatory frameworks** offering the technical and governance mechanisms to ensure that data use and sharing comply with EU regulations, values and objectives. Within the next part of the chapter, the multifaceted way that Data Spaces support GenAI is offered in more detail. Particular emphasis should be put on how the **investments and achievements in data spaces** form a reality (**foundations**) that can be leveraged on fostering **GenAI growth** and innovation.

Data Spaces: A data-sharing infrastructure addressing data needs of GenAI

- **Entry point to Domain specific (including industrial) data needs:** The need for case- or domain-specific data has been elaborated in chapter 2 for the development of specialised and efficient GenAI applications. These datasets are aimed at contributing to the adaptation of existing models, application integration or the creation of smaller models from scratch, tailored to the needs of each domain. Under their current set-up data spaces are operating under a sectorial/domain-specific focus (e.g., manufacturing data space, health data space, mobility data space), and include data products which offer access to even proprietary, industrial data (not accessible under other circumstances or environments). These could encompass – sometimes fresh and untapped – domain specific data, such as for example industrial machinery data from the manufacturing data space, and multilingual content from publishing houses included in the language data spaces, which could be available under contractual agreements and allow for further enhancements in model contextual performance.

Addressing: Data needs

- **Providing strong foundations for High Quality Data for GenAI models and applications:** While there are different data needs associated with the various stages of a GenAI application development (i.e., for training, adapting or augmenting, for example, an LLM) these developments nevertheless highlight the importance of large-scale high quality data sources and emphasise the need for data availability, discoverability, accessibility and interoperability. Amid concerns about, e.g., bias, errors, ethics, relevance, mode collapse, quality, etc., of the data used for GenAI (as mentioned in chapter 2), it is obvious that the quality and relevance of this data is very important for ensuring the ethical and trustworthy development of GenAI.[77] Data spaces can offer the foundations in terms of means, tools, processes to assess the quality of data (knowledge about data source, lineage and provenance – potentially also mitigating the risk of content pollution); can describe it properly (under exhaustive, reliable and standardised descriptions that require less preparations and foster accelerated model developments); and can ensure that the data quality is maintained during operations, while the whole process is governed properly (under governance rules and services).

Addressing: Data needs, other needs

[77] BDVA (2024), Elevating Data Quality A Paradigm Shift for Data Spaces and AI Needs, May 2024, version 1.0; CC BY-SA 4.0, available at <https://bdva.eu/news/new-paper-%F0%9F%91%89-elevating-data-quality-a-paradigm-shift-for-data-spaces-and-ai-needs>

- **Contributing to Privacy-Preserving GenAI through distributed approaches:**

Data spaces can enable participating organisations with a wide range of opportunities to train GenAI models cooperatively on dispersed datasets while maintaining data security and privacy. This distributed approach not only promotes the creation of reliable GenAI models but introduces further emerging opportunities such as Federated Learning (FL), where GenAI model training doesn't require data (including private data samples) to be moved to a central computing environment. This method ensures that data (considering also sensitive/personal datasets) remains at their source (ensuring security), lifting the need for unnecessary data movements (fostering eco-friendly practices), while still enabling GenAI advancements.

Addressing: Data needs, computing needs, other needs

Bringing up the digital (business) ecosystem perspective

- **Developing a community of stakeholders:** Data spaces are emerging at the core of industrial and business ecosystems as one of the core elements to share data in a trusted way, implementing a governance framework that establishes rules of collaboration, and helps to develop a community of different stakeholders. Considering this from the GenAI perspective, data spaces offer access to data holders (enabling sharing of domain specific datasets) and to data value creation service providers (offered by some of the data space participants) that can offer required services for GenAI (i.e. data preparation activities, Privacy Enhancing Technologies (PET) functionalities, evaluation frameworks for data, models, etc.). Data spaces, following interoperability principles and the use of common standards, lay the ground for cross-domain synergies (among sectoral data spaces) but also synergies with other environments (computing, AI initiatives), facilitating additionally the identification of common community needs (interoperability, skills, etc.)

Addressing: Business & Financial perspectives of GenAI, Skills needs, Data needs, Computing needs

- **Distributed value chains and business and monetisation opportunities:** Data spaces provide the groundwork for fostering partnerships and distributing value chains in the development of GenAI, ensuring a more fair distribution of value across stakeholders involved. By providing governance and a controlled environment, data spaces allow their participants to know which data to access, under what conditions and at what cost (industrial/proprietary data are not shared for free), allowing for diverse forms of data monetisation and contributions to the GenAI value chain. This creates favourable conditions for gathering large amounts of data, but also opportunities for a wider set of stakeholders – from large organisations to even smaller players that would find it impossible to participate individually - to form partnerships and engage in the various phases of GenAI projects (both on the model or application side and on the data providing side). This opens the door for more business & innovation opportunities:
 - Once a GenAI model has been established, the participants in a data space may wish to adapt or augment it with their data, and possibly with data from other participants of the data space. Such capability will be a constant feature of data spaces as GenAI continues to evolve allowing for further optimisations and new opportunities to innovate.
 - AI systems will be able to produce data that can be further exchanged, re-used and when relevant monetised/commercialised through the data space.

Addressing: Business & Financial perspectives of Generative AI, Data needs

Supporting Regulatory Compliance

Data spaces can play a crucial role in supporting the development of trustworthy GenAI by providing the necessary technical infrastructure and governance mechanisms to help ensure that data re-use and sharing align with the relevant EU regulatory framework, adhere to harmonised standards, and remain transparent and trustworthy. Data spaces, functioning also as a “space of rules”, are intended to provide secure environments for data storage and sharing, and for enhancing collaboration and innovation, while maintaining high levels of privacy and security, and fostering compliance with relevant EU legislation and acquis, including GDPR, copyright law and – most recently - the AI Act, etc. Data spaces are also designed to foster trust among participants but also offer the foundations for transparency and traceability in data handling practices.

- **Promoting transparency and traceability:** According to the AI Act, and as mentioned in section 2.4 of this whitepaper, GenAI (or GPAI) models must adhere to certain transparency requirements. For example, the AI Act mandates GPAI model providers to make publicly available a sufficiently detailed summary of the content used for training the model (Art53(1)(c)). [78] Data holders and data subjects whose data is made available through data spaces should have the possibility to be aware of, and to provide permission for their data and content being used to train these models. The role of data spaces in tracking and documenting data-related activities can enhance transparency and help ensure that AI providers remain accountable and comply with regulatory requirements throughout the GenAI lifecycle.

Addressing: transparency and accountability needs for Generative AI development

- **Facilitating adherence to standards:** By providing a common framework for data sharing, data spaces can support the harmonisation of technical and legal standards across sectors and prepare the way towards higher standardisation in data governance. They can enhance the FAIR (Findability, Accessibility, Interoperability and Reusability) usage of data[79] by encouraging the adoption of standard, compatible formats and protocols for coherent data gathering and processing, thus paving the way for generic tool development and aiding businesses that are active across multiple data realms. As described in case studies in Part 2, they can also assist in normalising the metadata required to identify problematic content by providing clear and complete information for each dataset, e.g. including licence information and dependencies between datasets.

Addressing: standardisation needs for Generative AI

- **Data governance:** data spaces can help facilitate compliance with data governance requirements (including those outlined in the EU AI Act and other relevant horizontal regulations such as the Data Act and the Data Governance Act, as well as relevant sectoral regulations). They can provide a foundation for access to quality, diverse datasets and to structured environments for data access and sharing, thereby serving as a basis for the development of trustworthy GenAI models and systems.

[78] Recital 107 of the AI Act offers some examples of how this summary can be achieved: listing the main data collections or sets that went into training the model, such as large private or public databases, or data archives, and providing a narrative explanation about other data sources used.

[79] FAIR refers to four foundational principles for data management and stewardship —Findability, Accessibility, Interoperability, and Reusability (FAIR). See Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1), 1-9.

This is particularly relevant in relation to compliance with obligations for high-risk AI systems or GPAI models in the AI Act. The below list outlines some specific benefits of data spaces in supporting data governance in correspondence with selected provisions of the AI Act, as well as with relevant horizontal requirements of that Act:

- Article 10 of the AI Act imposes strict obligations on AI providers regarding data governance and quality. Data spaces can enable the discovery and accessibility of well-curated, quality datasets, increasing the chances that data used in training, testing, and validation processes for GenAI is relevant, representative, and free from unfair biases. Additionally, they can provide standardised protocols for data collection and annotation, helping to ensure compliance with governance standards, including privacy preservation and trustworthiness, as mandated by Article 10(5).
- Data spaces can also streamline information-sharing obligations under the AI Act, potentially serving as platforms that can facilitate tracking of data provenance and where technical documentation and compliance records can be maintained. Providers of high-risk and general-purpose AI models (defined according to the AI Act) can ensure that data is shared along with associated technical documentation, thereby meeting the requirements of Articles 11, 12, and 53. This reduces friction in regulatory processes and enables seamless collaboration across the AI value chain.

Addressing: Data governance needs of Generative AI; compliance with AI Act and other relevant EU regulation.

- ***Promoting protection of intellectual property rights (IPR):*** Data spaces can contribute to ensuring that data sharing and processing comply with EU laws in this area, such as the relevant copyright framework[80] and the AI Act. For example, they can offer the following:
 - **Standardisation of Data Sharing Practices:** By creating standardised protocols and frameworks for data sharing, data spaces can help ensure data is shared in a way that valorises IPR. This includes clear licensing agreements and usage rights.

[80] e.g. Directive 2001/29/EC of the EP and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society [Info Soc Directive] and Directive (EU) 2019/790 of the EP and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC.

- **Secure Infrastructure:** Data spaces can provide a secure infrastructure with protection against unauthorised access, data breaches, and cyberattacks, which helps prevent theft and misuse of intellectual property.
- **Access Control:** Data spaces can help implement robust access control mechanisms, ensuring that only authorised entities can access certain data sets. This can help to prevent unauthorised distribution of copyrighted content and protect the rights of content creators and owners.
- **Traceability and Transparency:** Data spaces can enable traceability of data transactions, which helps establish a clear chain of custody for digital assets, making it easier to enforce copyright and intellectual property rights. Data spaces support more effective oversight and accountability in GenAI development by enforcing usage policies and enabling the tracking of datasets feeding a GenAI model.
- **Data Provenance and Attribution:** Data spaces can maintain records of data provenance and attribution, which can assist in proving ownership and originality of content, deterring plagiarism, and supporting the enforcement of copyright claims.
- **Respecting rightsholders rights to opt out from training of AI models:** Data spaces could offer an environment within which rightsholders and authors are able to adopt a more granular and substantive mechanism to opt out from the use of their works to train AI systems, for example allowing them to specify the purposes for which their works can be used or offering a streamlined environment for licensing their rights to AI developers.

Addressing: Copyright and other IPR needs of Generative AI; compliance with relevant EU regulation

- **Supporting compliance with privacy and data protection law:** By providing a common framework for data sharing, data spaces can ensure that all actors within the data space are adhering to the same data protection principles and requirements, which can simplify GDPR compliance for businesses and organisations. Data spaces are developed with a focus on security and privacy-by-design principles. They offer secure environments where data can be shared and processed with proper contracts and technical safeguards in place, such as encryption and anonymisation techniques, which are key tenets of GDPR compliance. Clear governance structures, consent mechanisms, and tracking of data provenance can contribute to ensuring that individuals' rights are respected, and that personal data is processed lawfully and fairly.

Data spaces can also facilitate the right to data portability, a right guaranteed by the GDPR, which allows individuals to obtain and reuse their personal data across different services - data spaces can provide the technical infrastructure needed to enable seamless data portability in a secure and compliant manner.

Addressing: privacy and data protection needs of Generative AI; EU regulatory compliance

- **Cybersecurity:** data spaces can assist in facilitating adherence to cybersecurity requirements in accordance with EU law (see sections 2.4.5 and Part 2: 3.2, the latter of which provides an example of how data spaces could help address cybersecurity challenges by creating secure environments where data can be shared and analysed without compromising privacy or security). This is an important area for further development in the future.

Addressing: cybersecurity needs of Generative AI; compliance with relevant EU regulation

Investments and key accomplishments of Data Spaces in EU as an enabler for GenAI

Over the past years, Europe has made significant investments to foster the deployment of data spaces and to support the European data strategy. These efforts, alongside key regulations (also mentioned previously in this chapter) aligned with EU values, have now established the foundations that can be leveraged to enable and drive developments for GenAI.

- **EU Investments in Data Spaces:** EU is actively funding several initiatives relevant to the deployment of the common European data spaces. These initiatives mainly fall under the following three categories: [81]
 - **Sector and domain specific data space (DS) initiatives** funded under the DEP and HE programmes: Including initiatives targeting to initiate, develop and maintain data spaces in 14 critical sectors/domains for the EU economy and specifically in: Agriculture, Cultural Heritage, Energy, Finance, Green Deal (including Smart Cities & Communities initiatives), Health, Language, Manufacturing, Media, Mobility, Public Administration (including Public Procurement) and Research & Innovation, Skills and Tourism. From preparatory to deployment actions, all sector and domain specific DS initiatives aim to ensure alignment and synergies for the effective realisation of the European strategy for data.

[81] <https://ec.europa.eu/newsroom/dae/redirection/document/101623>

- **The Data Space Support Centre (DSSC):** This is a virtual organisation and EU-funded project under the DEP which supports the deployment of common European data spaces and promotes the reuse of data across sectors.[82] The DSSC collaborates with a broad Network of Stakeholders (including sector and domain specific data space initiatives) to co-create and share assets covering different aspects related to data spaces. DSSC assets are important offerings towards supporting the deployment of data spaces.
- **Simpl:** A smart open-source & secure middleware platform that ensures data spaces will be deployed based on a common technical infrastructure. It supports data access and interoperability among data spaces. Simpl started in January 2024 and will develop the following products: Simpl-Open, Simpl-Labs, Simpl-Live.

Next to these initiatives, there are several R&I projects funded under the HE programme which further support the realisation of data spaces by building relevant components and technologies, and by contributing with building blocks across several verticals to help shape the future of data spaces (e.g. projects funded under HE Cluster 4 – 2021/2022 – Data – 01, or under HE Cluster 4 -2023-Data-01-02).

- **Relevant legislative instruments:** To implement the overall European data strategy vision several cross-sectoral legislative instruments have been adopted, with the aim of ensuring fairness in the digital environment, stimulating a competitive data market, opening up opportunities for data-driven innovation and making data more accessible for all. These include, inter alia, the **Data Governance Act**,[83] introducing a set of horizontal measures to boost trustworthy data sharing; the **Data Act** (which will take effect from 12 September 2025),[84] aiming to make more business and non-personal data available for reuse through the definition of rules on who can access and use what data and for which purposes; and the **Implementing Act on High Value Datasets**, implementing the Open Data Directive to facilitate the reuse of high value datasets in six thematic areas: geospatial, earth observation and environment, meteorological, statistics, companies and company ownership, and mobility. Specifically relating to public sector data, the Interoperable Europe Act also promotes data sharing, reuse and collaboration through an interoperable network of digital public administrations, thereby facilitating the creation of a Common European Public Sector Data Space. [85]

[82] See DSSC glossary.

[83] Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724.

[84] Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828.

[85] Regulation (EU) 2024/903 laying down measures for a high level of public sector interoperability across the Union (Interoperable Europe Act).

The background features a vibrant green field with large, overlapping, curved shapes in a bright blue color, creating a dynamic, organic pattern.

4 Challenges to unlock the full potential of data spaces for Generative AI

Data spaces offer substantial potential for a wide range of opportunities in GenAI. However, realising this potential requires overcoming several key challenges under the current setup of data spaces which may create barriers to the efficient and effective integration of data spaces and GenAI. In this chapter, we discuss about these specific challenges in detail.

1 Maturity of data spaces in comparison to the growing demands of GenAI:

The majority of the common European data spaces haven't reached the level of maturity needed to support the demands of GenAI at scale yet. Despite their potential and benefits towards GenAI, sectoral data spaces are progressively passing from preparatory to deployment actions, and the current landscape of data spaces is still diverse, complex and requires harmonisation. The fact that data spaces are not fully aligned with the stage of deployment and adoption needed for coping with present GenAI market needs (i.e. requiring to be in operational or scaling stage so that more and better data are available - both open and proprietary - to train models and foster developments) is reflected in the recommendation section as well, indicating its importance for unlocking the full potential of this interplay.

2 The lack of a holistic framework and integrated digital infrastructure to address the diverse needs and challenges of GenAI at scale:

While data spaces can offer concrete benefits for GenAI (as presented in section 3), there are still many challenges and needs (including data related ones) that require integration of several digital infrastructures and technologies to support the growth of GenAI. The main challenge is on one hand the lack of a common framework considering regulatory compliance, data management, computing resources, infrastructure, business models, and skill requirements all together, rather than addressing them as separate issues.

On the other hand, the challenges include the lack of maturity in integrating data spaces with other existing or emerging digital ecosystems and platforms, which can offer support in critical areas like data preparation, maintaining data quality, and adhering to industry standards and therefore enable seamless data preparation, cataloguing, and AI model accessibility at scale. The challenge also involves bringing all these elements together in a way that's flexible and scalable while fostering cooperation and innovation across different industries, a variety of stakeholders and sectors. Digital infrastructures for data sharing should not only serve large organisations but also enable smaller businesses across Europe to benefit equally. The computing continuum must be designed to reach downstream applications, enabling smaller participants to access the necessary tools and resources to drive AI innovation in a sovereign and competitive manner. Without inclusive and scalable infrastructure, there is a risk that smaller actors may be left behind, stifling broader AI-driven advancements across the European ecosystem.

- 3 Interoperability and simplicity:** As per today access to data through data spaces lacks simplified, harmonised access to cross-sectoral and federated data spaces. For stakeholders such as startups, SMEs, and initiatives contributing to European AI infrastructure (e.g., AI Factories), it is impractical to connect with every individual data space to meet their training data needs. Ad-hoc solutions to data access hinder efficiency, making it crucial to establish a unified approach for accessing these diverse and distributed data resources.
- 4 Data representation, diversity, and quality:** for GenAI models to perform effectively across a variety of scenarios (both within a specific context and while generalising), the datasets accessible via data spaces must be diverse, high-quality, balanced and represent a wide range of domains, topics, contexts and modalities. Ensuring that models can adapt to diverse inputs requires high-quality data that should capture a broad spectrum of variations (including different modalities, linguistic variations, etc), requiring rigorous data cleaning, validation processes, and continuous monitoring to maintain the integrity and usefulness of the datasets in data spaces.

The challenge here relates to: 1) the need to scale adoption of users in common European data spaces enlarging the possibilities to access a variety of data sources; and 2) there are complexities related to the design and management of data spaces that must be addressed to meet GenAI requirements. Developing design methodologies, approaches, and components that reflect the needs of GenAI—such as data models, provenance, traceability, data quality, governance, and ethics—is essential to prevent further fragmentation. Ensuring these aspects are incorporated into existing data space assets and frameworks presents a significant challenge, as failure to do so could lead to inefficiencies, inconsistencies, and obstacles in leveraging data spaces for GenAI development.

5 Lack of sufficient evidence or examples on how successful GenAI applications have been built based on data spaces/sharing paradigms.

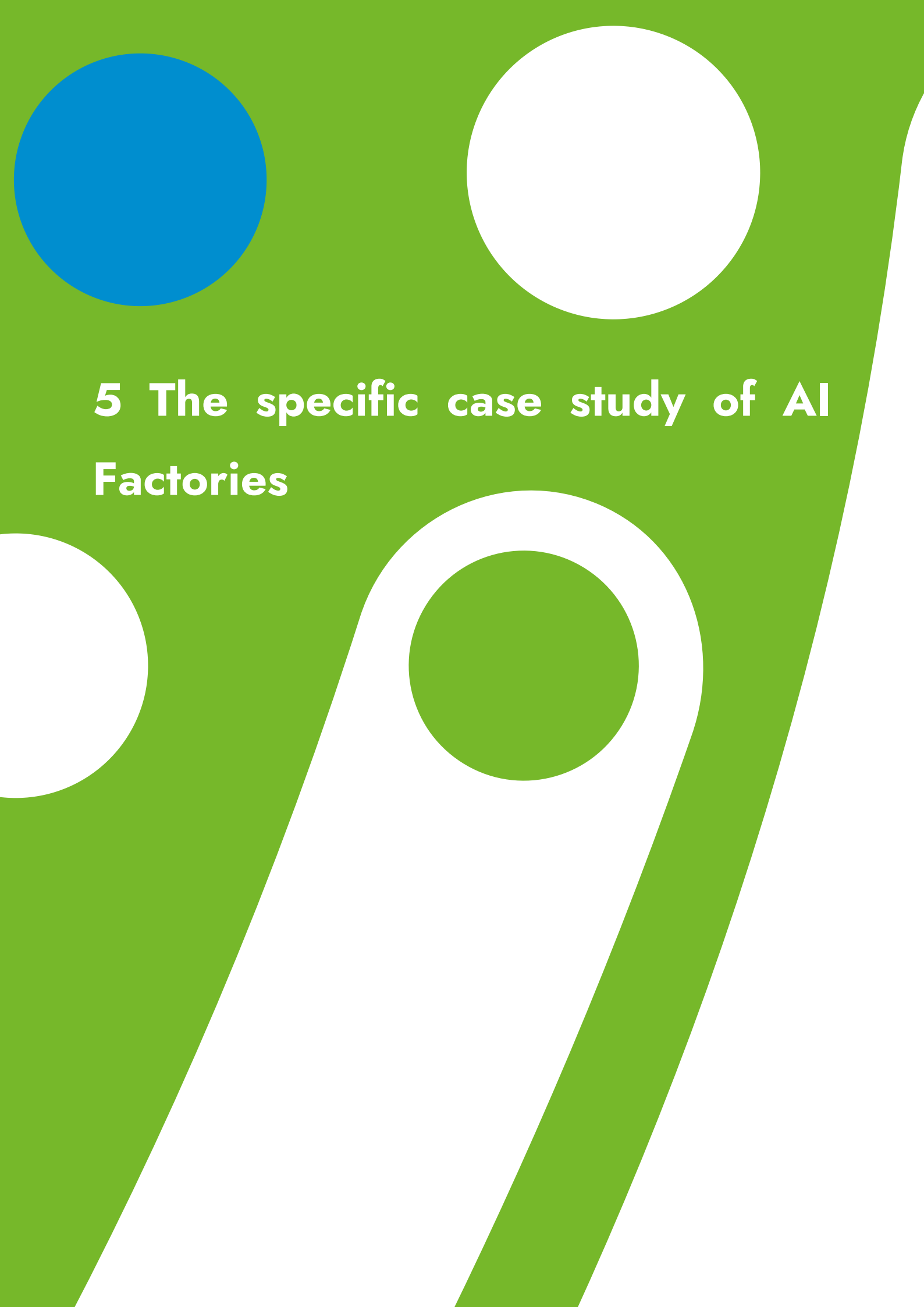
Another challenge for data spaces, more related to business and adoption aspects, is the absence of examples and success stories that demonstrate how GenAI applications have been successfully built in a context of industrial/private data sharing ecosystems. This lack of evidence creates difficulties in showcasing the tangible benefits of the interplay among data spaces and GenAI and consequently the business opportunities available to new market players. The difficulty lies in gathering diverse, domain specific and cross-sector use cases that showcase how GenAI can drive value within data spaces while also guiding new entrants in identifying potential business models, partnerships, and roles they could assume.

6 Lack of incentives to leverage Industrial data for GenAI in data spaces:

Another challenge in utilising industrial data for GenAI applications within data spaces is ensuring that this potentially underutilised resource is effectively accessed and exploited. While industrial data presents a valuable opportunity for training, adapting, fine-tuning AI models and creating and using applications, and new business opportunities, convincing participants of these benefits is essential to driving broader adoption of data spaces. Thus, managing the expectations of participants and aligning their business models with the goals of GenAI development remains a critical obstacle.

The availability of rich industrial data may be limited by established contractual agreements, which may be part of the business models of participating entities. Ensuring that these agreements allow for transparent and compliant data sharing is essential but can be complex. Furthermore, another challenge is ensuring the provenance, quality, and transparency of the industrial data used for AI training (as reflected in challenge 4). Data spaces must ensure that these factors are addressed to maintain trust and reliability in AI outputs.

- 7 Adaptation to new data trends:** data spaces must remain attuned to emerging data trends and shifts in language use. Mechanisms for quickly incorporating new datasets that reflect current events, societal changes, or evolving language patterns are essential. This adaptability ensures that the generative models trained on these datasets stay relevant and effective over time.

The background features a vibrant green field with white, wavy, ribbon-like shapes that curve across the frame. Several circles are scattered throughout: a solid blue circle in the top left, a solid white circle in the top right, a solid white circle on the left side, and a green circle with a white border in the center-right.

5 The specific case study of AI Factories

Introduced by the European Commission as part of the AI Innovation Package, AI Factories provide an illustrative example where data (combined with computer power and talent) is supposed to fuel the potential of GenAI and emerging AI applications in Europe. They also present a scenario where most of the aspects and learnings presented in previous sections of this White paper can be applied.

The content of this section is based on an ongoing discussion paper that BDVA is currently elaborating with the purpose of providing coherence and collecting common knowledge on the topic of AI Factories, and to propose recommendations and a roadmap of actions to accelerate access of the AI Factories to the Common European data spaces. This paper will be published in the upcoming weeks and made available on the [BDVA webpage](#).^[86]

AI Factories are open ecosystems formed around European public supercomputers and bringing together key material and human resources needed for the development of GenAI models and applications.^[87] AI Factories are created with the objective of leveraging those EU supercomputers for the development, training and deployment of large AI models, to support AI startups and small and medium-sized enterprises (SMEs), science and innovation ecosystems, and other productive environments. While AI Factories are especially targeted towards startups and SMEs, their activities will be open to all public and private users.

It is expected that a number of AI Factories will be established across Member States to serve European and national AI communities, therefore every AI Factory needs to be linked with the national AI strategies and data policies, as well as linked to the selected AI ecosystems.

[86]<https://bdva.eu/downloads/>

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

AI Factories are expected to include computing power, data and supercomputing services as well as large-scale talent attraction activities:

- AI Factories are centered around European AI supercomputers.
- As data is the fuel for AI development, it is also central for AI Factories to put in place measures that ensure access to the datasets that are relevant in the targeted AI ecosystems.
- AI Factories should identify communities with the required AI talent available and support upskilling of people.

Activities of an AI Factory include:

- Provision of AI supercomputer resources according to AI-friendly access policy.
- Provision of data storage resources.
- Provision of high-bandwidth connectivity.
- Enabling access to datasets and Common European data spaces and complementary data repositories.
- Provision of rich software environment for AI.
- User support.
- Provision of co-working and hub facilities.
- Training and education.
- Engagement with AI communities.
- Networking with other AI Factories.
- Developing Trustworthy AI.

To achieve their objective, AI Factories should be able to access, acquire and reuse large amounts of different types of data coming from different sources, repositories and data ecosystems. Different access, use and re-use terms and conditions, and specific regulations may apply to different datasets and data ecosystems, particularly important for privately owned data and personal data.

EU data spaces emerge as ideal instruments to provide AI Factories with seamless access to data coming from a diversity of sources and under different conditions. Data spaces can also support AI Factories in dealing with data interoperability, security, governance and quality.

However, the connection between AI Factories and data spaces also poses many challenges at different levels that should be addressed in order to guarantee smooth access by AI Factories to the data required to develop and train large AI models.

A key aspect to take into account is the timing alignment between AI Factories and data spaces, considering their different level of maturity. At present, the AI Factories are being discussed with Member States, and potential hosting organisations are defining their strategies. The operation of the first AI Factories is expected soon - with notification of results of the calls launched for AI Factories expected in December 2024 and tight deadlines also indicated within the EC President's recent mission letter^[88] to the Executive Vice-President-designate for Tech, Sovereignty, Security and Democracy. At the same time, the overall EU data space deployment programme is underway, setting up the main European infrastructures for secure, sovereign and trustworthy data sharing. So in many aspects, AI Factories can build on the work already done within European data spaces, which would require a certain level of alignment.

Among others, some challenges to be considered are:

- Availability of high amounts of quality data relevant for the purposes of users of AI Factories (SMEs and especially start-ups). Data spaces have to be sure that they are able to provide, in a timely manner, the constantly growing large datasets that the training of large-scale foundation models requires.
- Interoperability at different levels. This should be addressed considering both existing international standards and upcoming ones, and aligned with the different ongoing efforts.
- Data and AI governance, including a collaborative scheme (according to capabilities of the data ecosystem and the lifecycle of the data and models) to ensure proper governance of data and AI (e.g., data sovereignty, the conditions for the reuse of data, data traceability), and compliance assurance (e.g., with GDPR, DGA, AI Act).
- Identity and role management. A common approach should be agreed between common European data spaces and AI Factories to register and identify users to determine access rights.

- Automation and scalability. Be able to automate as much as possible 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.
- Integrate into the business model of the AI Factories the new conditions and costs associated with the process of accessing, acquiring and reusing data (fees to subscribe to data ecosystems and repositories, price to pay to data sources and / or data providers according to initial policies and conditions attached to the datasets, and potentially elaborated further in contracts negotiation).
- Lack of appropriate skills. AI Factories need access to specific skills related to the acquisition, management and governance of data (including Data Protection Officers) in order to be able to train stakeholders and support end users. For end users, and depending on the level of abstraction and support provided by the AI Factory, skills are required to manage and use data at the user level and according to its final purpose.

Taking all the above into account, the realisation of the integration between AI factories and data spaces requires establishing a well-defined roadmap of activities. HPC ecosystems must evolve from traditional isolated scenarios to be able to align and connect with other digital ecosystems (including data spaces). It would also be desirable to have a certain level of harmonisation in the connection of AI Factories to data spaces and access to data, to avoid ad-hoc solutions for each AI Factory connecting to specific data ecosystems. AI Factories and data spaces should also be able to converge in innovative business models so they can benefit from this integration and from each other. Finally, to achieve this objective, AI Factories would also benefit from their connection to European sandboxes and data experimentation environments (e.g. BDVA i-Spaces).

An abstract network diagram with blue lines and circles on a green background. The diagram features several blue circles of varying sizes connected by blue lines. A large circle is in the top right, with lines radiating from it. Another large circle is in the bottom left, and a medium-sized circle is in the bottom right. A small circle is in the center. Lines connect these circles in a network pattern, with some lines extending towards the edges of the frame.

6 Recommendations

As a result of the discussions that emerged during the creation of this paper, and the outcomes of some of its chapters, a list of recommendations is proposed. The aim of these recommendations is to prioritise areas of common work by the addressed stakeholders and the involved communities, thus, influencing future developments in the interplay between GenAI and data spaces, fostering growth and synergies between the involved ecosystems.

The recommendations target different audiences. Some of the recommendations are relevant for the data space initiatives and the DSSC while some others could be more transversal and relevant for a broadened set of stakeholder groups: including policy makers, existing industrial ecosystems, European GenAI infrastructure initiatives, data space initiatives and the DSSC and to a lesser extent GenAI providers (i.e. General purpose AI model providers and downstream providers across multiple sectors).

Recommendations more relevant for data space initiatives and the DSSC:

→ **Define and deploy means of simplified, harmonised access to cross-sectoral and federated data spaces serving data accessibility for GenAI needs.** For interested stakeholders (including individual entities involved in GenAI like startups and SMEs or initiatives contributing to European GenAI infrastructure like AI Factories) it would be impractical to be connected to multiple data spaces in order to address their training data needs, following ad-hoc approaches. Solutions spanning from “lightweight” connections as simple as being a data space participant (or a “special” participant in the case of initiatives like AI Factories connecting to data spaces) to full interaction with mechanisms and conditions to scale-up the access and reuse of data need to be considered. Data spaces should consider this recommendation for lifting barriers relevant to interoperability and cross-sectoral data sharing.

The development and adoption of international standards[89] plays a crucial role in this recommendation. This recommendation mainly targets the data space community at large (data space initiatives, participants, DSSC, standardisation organisations and policy makers).

→ **Foster innovation in data ecosystems with GenAI.** Data spaces could enable and foster innovation, allowing their participants to experiment and unlock the full potential of GenAI. With GenAI sandboxes or experimentation environments built in each common European data space, interested stakeholders could start experimenting with possibilities that each ecosystem gives to different players in the value chain of GenAI. Sandboxes will help the creation of innovative projects, enabling the identification of important missing pieces but also of further necessary recommendations (i.e. for data quality improvements, adaptation to emerging data trends) to further encourage breakthroughs. This recommendation mainly targets the DSSC and data space initiatives.

→ **Identification of use cases and creation of a catalogue of learnings from GenAI applications built on top of data sharing – using it as an instrument to further promote data sharing:** As data spaces become more and more mature and scenarios of data spaces contributing to GenAI models and application advancements emerge naturally this recommendation suggests creating a catalogue of use cases and learnings across different sectors that highlight insights gathered, challenges encountered and benefits of data sharing in GenAI. Such a catalogue would serve as a valuable resource for newcomers and organisations eager to support GenAI or data space developments offering concrete incentives and demonstrating that the synergy between GenAI and Data Spaces holds a significant transformative potential. This recommendation mainly targets the DSSC and data space initiatives.

[89] Aligned with different ongoing efforts (i.e. EC data standardisation request on data and data spaces, CEN CENELEC Trusted Data Transactions, CEN CENELEC Focus Group Data, Dataspaces, Cloud and Edge)

→ **Advance in integrating GenAI requirements in data space assets, frameworks, technologies and participants.** Data spaces should explore if current assets and building block descriptions (i.e. data models, provenance & traceability, governance, data, services and offerings descriptions, etc.), technologies and frameworks are considering requirements for GenAI developments and the support of relevant EU enabling infrastructures (i.e. AI Factories) in a sufficient way. For addressing GenAI requirements and challenges, data spaces should consider emphasising further on aspects including data quality, diversity & representation (also aiming for clearly labelling or segregating AI-generated content for preventing inadvertent pollution of training datasets) or sustainable data operations and management (i.e. already reflected in the design principles). Moreover, data spaces could consider investing (most likely through the value creation building block) in scalable storage infrastructure, high-throughput data processing pipelines, and automated data curation tools to handle the data influx required for training GenAI models. Some of these elements could be provided by data space participants. The engagement with relevant initiatives and stakeholders in the European GenAI ecosystem should also be prioritised within the evolution of the DSSC Network of Stakeholders to enhance collaboration and innovation. This recommendation mainly targets the DSSC and data space initiatives.

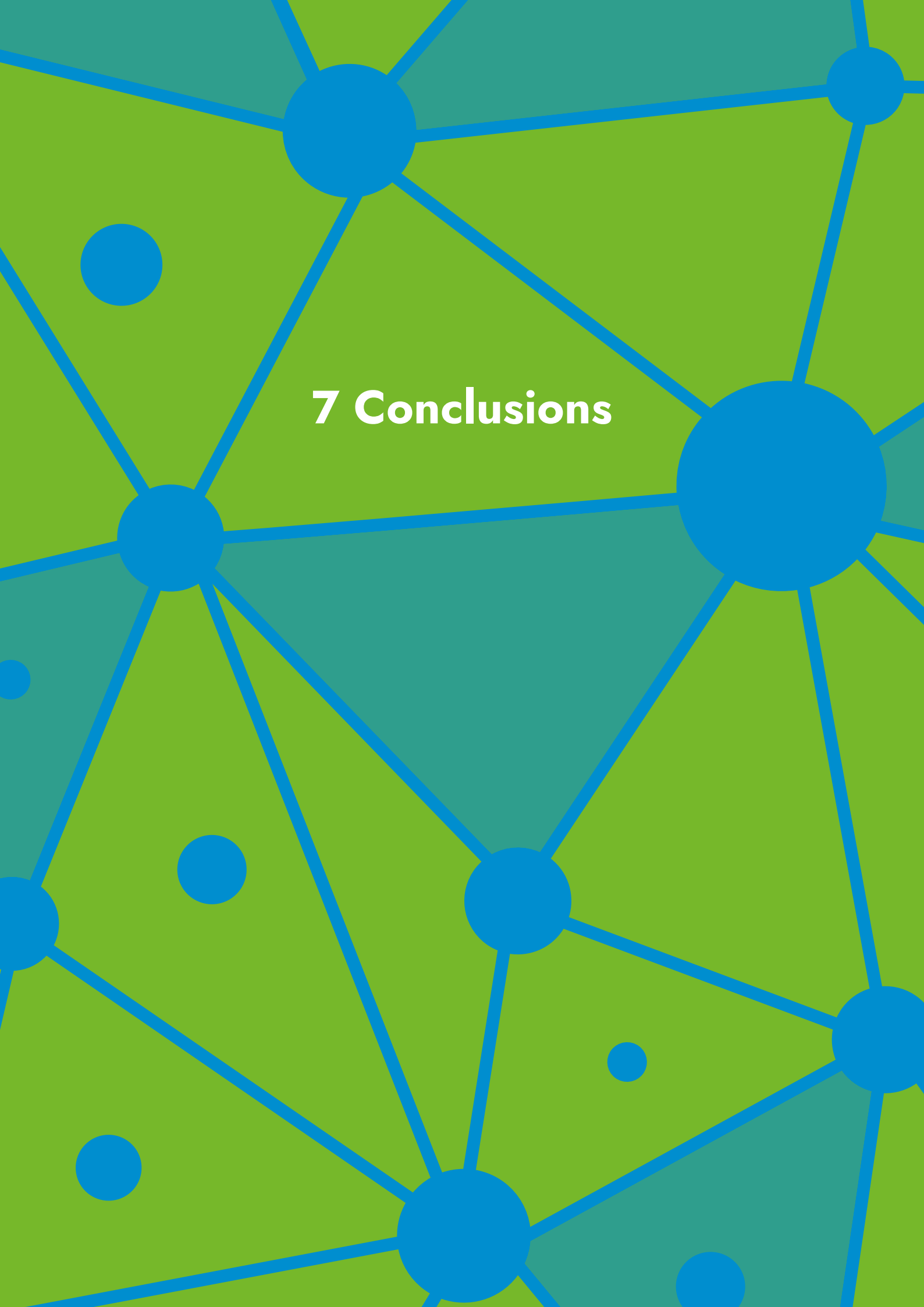
More transversal recommendations:

→ **Coordinate efforts for creating domain-specific GenAI models across all industrial sectors by building on common European data spaces:** Including models, either foundational or adapted/fine-tuned covering the needs for a specific domain and/or industry ecosystem and offering a plethora of opportunities including tailored interactions, operation efficiencies and hallucination reductions. As the demand for more niche solutions increases, the need for highly reliable, domain-specific models is also expected to increase. By adopting this recommendation, stakeholders gain access to reliable, domain-specific and data space-aligned GenAI models, creating new business opportunities & sectorial benefits, lowering entry barriers, and ensuring the reliability and compliance of generated outputs with established industry standards. This recommendation requires collaboration of all the above-mentioned stakeholders.

→ **Continue balanced and coordinated investments in data spaces, GenAI, and Computing infrastructures.** Shaping a balanced distribution of efforts across GenAI - Computing - Data is necessary to ensure growth in all fronts and promotion of synergies between the respective ecosystems (sectorial and cross-sectorial). This recommendation is mainly targeting policy makers and industrial players.

→ **Shape an integrated & comprehensive framework to address the diverse needs of GenAI in a holistic and timely manner.** This framework should consider key areas like regulatory compliance, but also address GenAI needs relating to data, computing, infrastructure, innovative business models and skills, not in isolation but rather combined. This recommendation aspires to intensify the integration and synergies of data spaces with other digital infrastructures relevant for GenAI, including HPC, quantum computing, the computing continuum, AlonDemand, and AI TEFs creating win-win scenarios. This recommendation requires collaboration of all the above-mentioned stakeholders.

→ **Accelerate training and upskilling:** New skills and competences are needed to deal with evolving requirements in data spaces, computing and GenAI and their intersections. This derives from both elaborating on skill needs from the GenAI perspective and very particularly on the more urgent needs on AI factories, where also lack of appropriate skills has been noted. Offering incentives and opportunities for upskilling the workforce in areas including: data productisation, management, governance and compliance, GenAI model development and deployment (different skills are needed per different stage of development, i.e. from legal & technical competences to human-centred and business skills), legal & ethical use of AI, cloud computing & architecture will be beneficial and create favourable conditions for exploiting the emerging opportunities. This recommendation mainly targets policy makers, academia, and industry at large.



7 Conclusions

This whitepaper is the outcome of a collective effort conducted with a broad set of stakeholders, part of the DSSC Network of Stakeholders. The value of the whitepaper is not only in the outcomes, but also in the co-creation and co-learning processes during its production. It has been built on the assumption that fostering a symbiotic relationship between data spaces and GenAI holds significant transformative potential for Europe. The focus point is on how data spaces can act as key enabler for trustworthy GenAI growth.

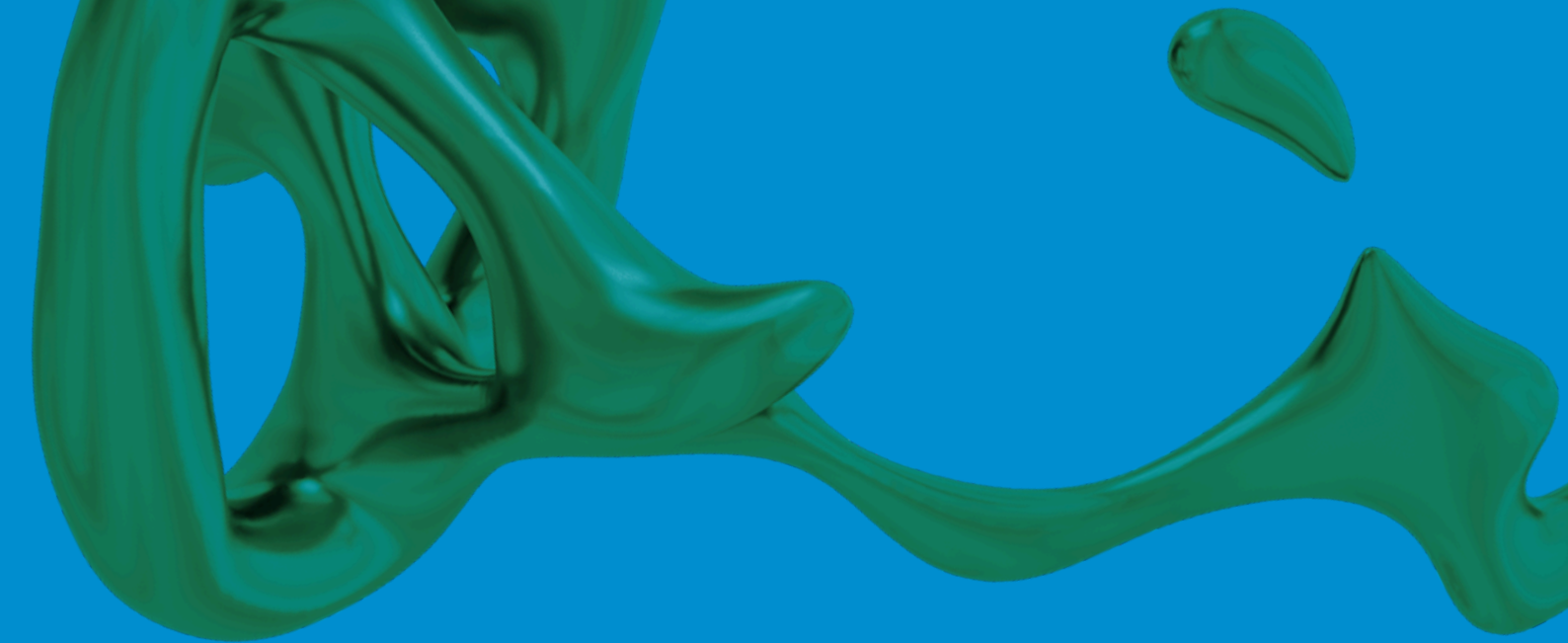
After assessing challenges and needs of GenAI in an holistic manner, the paper describes how data spaces address some of them and bring new opportunities: Data spaces provide the foundation for secure and efficient data sharing enabling access to a wide range of domain-specific datasets, including those from industrial sectors. Additionally, data spaces promote a collaborative digital ecosystem, driving innovation and creating new business opportunities for stakeholders. Data spaces also play a critical role in ensuring compliance with legal and regulatory frameworks, offering the tools needed to align data sharing with EU regulations, values, and objectives.

However to unlock the potential of data spaces for GenAI we still need to overcome many challenges and barriers including limited maturity, a lack of a comprehensive holistic framework for GenAI, and insufficient integrated infrastructure. Key issues include the need for greater interoperability, data diversity, and quality, along with simplified processes. Additionally, there is a shortage of evidence demonstrating successful GenAI applications based on data sharing, a lack of incentives to leverage industrial data, and a need to adapt to emerging data trends.

A set of 8 recommendations is proposed to a variety of stakeholders:

- 1) Coordinate efforts for creating domain-specific GenAI models across all industrial sectors by building on common European data spaces;
- 2) Continue balanced and coordinated investments in data spaces, GenAI, and Computing infrastructures;
- 3) Shape an integrated & comprehensive framework to address the diverse needs of GenAI in a holistic and timely manner;
- 4) Define and deploy means of simplified, harmonised access to cross-sectoral and federated data spaces serving data accessibility for GenAI needs.
- 5) Foster innovation in data ecosystems with GenAI;
- 6) Create a catalogue of use cases and learnings from GenAI applications built on top of data sharing paradigms;
- 7) Advance integration of GenAI requirements in data space assets, frameworks, technologies and participants;
- 8) Accelerate training and upskilling

While the momentum surrounding this interplay is promising; it is still understood that sustained, coordinated efforts are essential to overcoming challenges, maintaining alignment, and driving further innovations from a sectoral and cross-sectoral point of view. It is crucial to continue bridging knowledge gaps, addressing evolving needs in GenAI with data spaces and following through on the recommendations laid out in this paper to help in realising the full transformative potential of this symbiotic relationship.



PART 2: Learnings from several initiatives



The ambition of part 2 of this whitepaper is to delve into concrete examples and use cases of GenAI across various domains. Consequently, to highlight the challenges faced, concrete needs and requirements identified, alongside how data spaces help address particular issues and offer tangible solutions. The level of granularity varies across use cases, reflecting the diverse nature of the challenges and solutions explored. This section is divided into three sub-parts each offering a distinct perspective: from learnings captured by initiatives working on the deployment of common European data spaces, continuing with the perspective from the GenAI user or model provider before concluding with learnings from use cases and examples coming from research institutions and other projects.

1 Learnings from some common European data space initiatives

1.1 Media domain: Trusted European Media Data Space (TEMS)

Trusted European Media Data Space (TEMS)[90], funded by the Digital Europe Programme (DEP), is the flagship European initiative to build a resilient, data-driven ecosystem in the media sector. GenAI, with its rapid development and high impact on the creative industries, has quickly established itself as a significant part of modern media and, thus, also TEMS.

AI technologies have started to significantly enhance media production and consumption efficiency, creativity, and personalisation. Automated content generation tools powered by AI algorithms streamline the creation of written articles, videos, and audio content. Media companies are now racing to harness the power of these latest advancements in AI technology to transform content creation, distribution, and consumption.

There are clear benefits, and value is captured from large volumes of data present within media partners involved in TEMS to train GenAI models. **The concern of these partners** is that their data is used without their permission (e.g. through data “crawling” on the Web) **to generate significant value – value for which media companies do not seem to get compensated**. Currently the typical situation with LLMs is that the outputs of creative industries are used in training the models without enough consideration on fair compensation for the content providers.

[90] <https://tems-dataspace.eu/>

It is the view of TEMS that this imbalance can be resolved with data spaces. On one hand, they can help protect contents from illegal use by GenAI algorithms or any other AI operator. Additionally, they can offer business models for media companies to benefit from AI companies and big online platforms looking to acquire their contents to enrich their models. On the other hand, data spaces offer AI companies a secured and trusted method to find relevant data for their LLMs in a safe contractual environment, making it easier for them to comply with transparency requirements of regulation like the AI Act.[91] TEMS has been in active conversation on these topics with leading European GenAI companies like Mistral.[92]

It is the quality and multimodality (text, images, 3D models, etc.) of data the media sector provides that offers a competitive edge for AI models. Especially with potential collaborations under negotiation with other European data spaces like LDS, TEMS would be able to provide a plethora of unique datasets for training, fine-tuning and augmenting AI models. It is now up to the TEMS consortium, consisting of large European media companies as well as technology providers specialised in data ecosystems, to develop the business models, legal framework and technical infrastructure to enable this.

[91] AI Act Recital 107

[92] <https://mistral.ai/>

1.2 Language Data: Common European Language Data Space (LDS)

The Common European Language Data Space (LDS)[93] is one of the official EU data spaces, the project is funded through a procurement contract in the Digital Europe Programme (DEP).[94] The development of a data space for language data has recently become increasingly relevant due to the importance of large language models (LLMs) and GenAI solutions for industrial and academic applications and their dependence on vast amounts of language data for training models, especially when considering the increasing push of the EU towards digital sovereignty. LDS is one crucial part of a bigger initiative towards more European independence and increased European participation in the global LLM and GenAI landscape (including research, development, innovation, application, deployment and also monetisation). The LDS consortium consists of private/public and commercial organisations with a track record in language technologies research and development. The project itself not only develops the technical infrastructure of the LDS (which itself is based on the Eclipse Dataspace Components, EDC) but it also develops the multi-stakeholder governance structure, several proof-of-deployment projects and it organises dozens of events towards promotion and uptake.

LLMs, including GenAI models, are trained on vast amounts of data. With regard to LDS, we can distinguish three broad classes of data: (1) research data (corpora etc.) – these are well-known annotated or raw linguistic corpora, compiled and made available by the NLP community; this class of data is well-known to the LDS consortium, all relevant European repositories and platforms such as European Language Grid (ELG)[95] will be attached and made interoperable with LDS;

[93] <https://language-data-space.ec.europa.eu>

[94] Georg Rehm, Stelios Piperidis, Khalid Choukri, Andrejs Vasiljevs, Katrin Marheinecke, Victoria Arranz, Aivars Bērziņš, Miltos Deligiannis, Dimitrios Galanis, Maria Gavrilidou, Maria Giagkou, Katerina Gkirtzou, Dimitris Gkoumas, Annika Grützner-Zahn, Athanasia Kolovou, Penny Labropoulou, Andis Lagzdīņš, Elena Leitner, Valérie Mapelli, Héléne Mazo, Simon Ostermann, Stefania Racioppa, Mickaël Rigault, and Leon Voukoutis. Common European Language Data Space. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakti Sakriani, and Nianwen Xue, editors, Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 3579-3586, Turino, Italy, 2024. European Language Resources Association (ELRA) and International Committee on Computational Linguistics (ICCL). May 20-25, 2024

[95] Georg Rehm, editor. European Language Grid: A Language Technology Platform for Multilingual Europe. Cognitive Technologies. Springer, Cham, Switzerland, 2023

(2) very large web crawls such as Common Crawl (including processed versions of Common Crawl) or the OpenWebSearch crawl are indispensable when it comes to developing LLMs and GenAI solutions; these crawls are necessary and well-known, they will also be integrated into LDS; (3) the most relevant class of data is large amounts of new data, especially from industry and novel data providers that has been, so far, inaccessible to European LLM/GenAI research and development. This is, for example, data from public or private broadcasters, data from libraries or data from information companies.[96] It is exactly this class of data that has the potential of making a difference in and for Europe since this class of data can play the role of a competitive advantage in the global race for better performing LLMs and GenAI solutions, especially those that are adapted to European languages and cultures, fully respecting European values and European regulation etc.

While the LDS user group, i.e., the group of organisations using the Language Data Space, is still under development (just like the LDS infrastructure itself), the goal of the LDS initiative is to assemble a very large number of especially novel data providers (belonging to the third class above) in the LDS user group so that the language data available in and through LDS can serve as an enabler and booster of the European LLM and GenAI landscape, including the overarching goal of providing language technologies to speakers of all European languages including official EU languages as well as co-official, regional and minority languages.[97]

At the same time, there is enormous potential with regard to using those LLM and GenAI applications that are developed with data available through LDS, within and for the LDS itself. This can relate to, for example, making available large amounts of synthetic data (i.e., generated data) for model training, using LLMs for filtering and curating language data, translating or otherwise processing or enriching data using LLMs, anonymising data using LLMs etc.

[96] Especially the use of broadcast data in LDS is explored in the following technical feasibility study for the European Parliament in more detail: Georg Rehm, Muhammad Zeshan Afzal, Aliki Anagnostopoulou, Mahta Bakshizadeh, Stephan Baumann, Cristina España-Bonet, Carlos Franzreb, Josef van Genabith, Annika Grützner-Zahn, Stefanie Hegele, Tim Herzig, Jakob Karolus, Siting Liang, Paul Lukowicz, Katrin Marheinecke, Julián Moreno-Schneider, Malte Ostendorff, Simon Ostermann, Alain Pagani, Tim Polzehl, and Sven Schmeier. European streaming platform for national news accessible in all EU languages: Technical feasibility study. European Parliament, European Parliamentary Research Service, 6 2023. STUDY Panel for the Future of Science and Technology. Scientific Foresight Unit (STOA) PE 740.249 – June 2023

[97] Georg Rehm and Andy Way, editors. European Language Equality: A Strategic Agenda for Digital Language Equality. Cognitive Technologies. Springer, 2023

1.3 Green Deal domain: Common European Green Deal Data Space (GREAT)

The Green Deal Data Space preparatory action GREAT project[98] was funded by DEP (Sept 22 – April 24). Its implementation action will follow, starting from 2025. The GDDS aligns with the EU Green Deal strategic actions in its commitment to a net zero society by 2050. The data needed to achieve this target expands to all societal and economic sectors.

GenAI provides the means to analyse and make sense of vast amounts and diversity of environmental data needed to address the disruptive transformation towards a sustainable society. Sustainable economic activities will need to comply with environmental objectives, such as those defined in the EU Taxonomy Regulation:[99]

- 1)climate change mitigation
- 2)climate change adaptation
- 3)sustainable use and protection of water and marine resources
- 4)transition to a circular economy
- 5)pollution prevention and control; and
- 6)protection and restoration of biodiversity and ecosystems.

The application of GenAI to accelerate the path to sustainability targets spans a wide field of opportunities that are emerging and evolving at a very rapid pace. For example, businesses will benefit from the handling of large amounts of data such as those relating to energy and natural resources consumption, waste management, scope of the greenhouse gas emissions and impact on biodiversity, in order to comply with regulatory requirements and automate the generation of sustainability reports, as requested by the CSRD regulation.[100]

Predictive ML models have made significant advances to become competitive and economical alternatives to tackle climate or weather predictions compared to traditional and physics based numerical weather prediction models.[101] The opportunities to accelerate advances in the transition from data to insights can be further reinforced by GenAI.

[98] <https://www.greatproject.eu/>

[99] <https://eur-lex.europa.eu/eli/reg/2020/852/oj>

[100] <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022L2464>

[101] <https://www.ecmwf.int/en/about/media-centre/science-blog/2023/rise-machine-learning-weather-forecasting>

Tropical Cyclone localisation and Tracking, a real use case

Tropical cyclones (TCs) are considered some of the most devastating natural events. On average, about 90 TCs form each year over tropical oceans worldwide, and global warming is intensifying their strength and destructive potential. As a result, accurately locating and tracking these cyclones has become a significant and intriguing area of research in weather and climate science. Historically, TCs have been detected using deterministic tracking methods that depend on subjective thresholds. An ensemble approach proposed by Accarino et al. (2023)[102] exploits multiple ML models for improving the localisation of TC centres. A deterministic tracking scheme is also defined on top of previously localised TC centres to reconstruct potential TC trajectories. The accuracy of the ML model results validation, and quality of the training data are therefore paramount in the provision of a trustworthy AI/ML cycle that can ultimately be used to protect lives and infrastructure damage. The pipeline proposed by Accarino et al. (2023), is composed by different manual steps:

1. Data collection: TCs tracks from IBTrACS archive[103] and ERA5[104] reanalysis data from the Copernicus Climate Change Service;[105]
2. ML models training and fine-tuning: to train a set of ML models, each one having different structural characteristics or hyperparameters;
3. Inference: to detect TCs in ERA5 data;
4. Results validation: to validate the quality of the approach and reduce false negatives.

Figure 3 illustrates how the combination of LLMs and multiagent systems could support a pipeline composed by different disconnected steps like the one proposed by Accarino et al. (2023) leading to modular, efficient and less error prone pipelines.

[102] <https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023EA003106>

[103] <https://www.ncei.noaa.gov/products/international-best-track-archive>

[104] <https://www.ecmwf.int/en/forecasts/dataset/ecmwf-reanalysis-v5>

[105] <http://climate.copernicus.eu/>

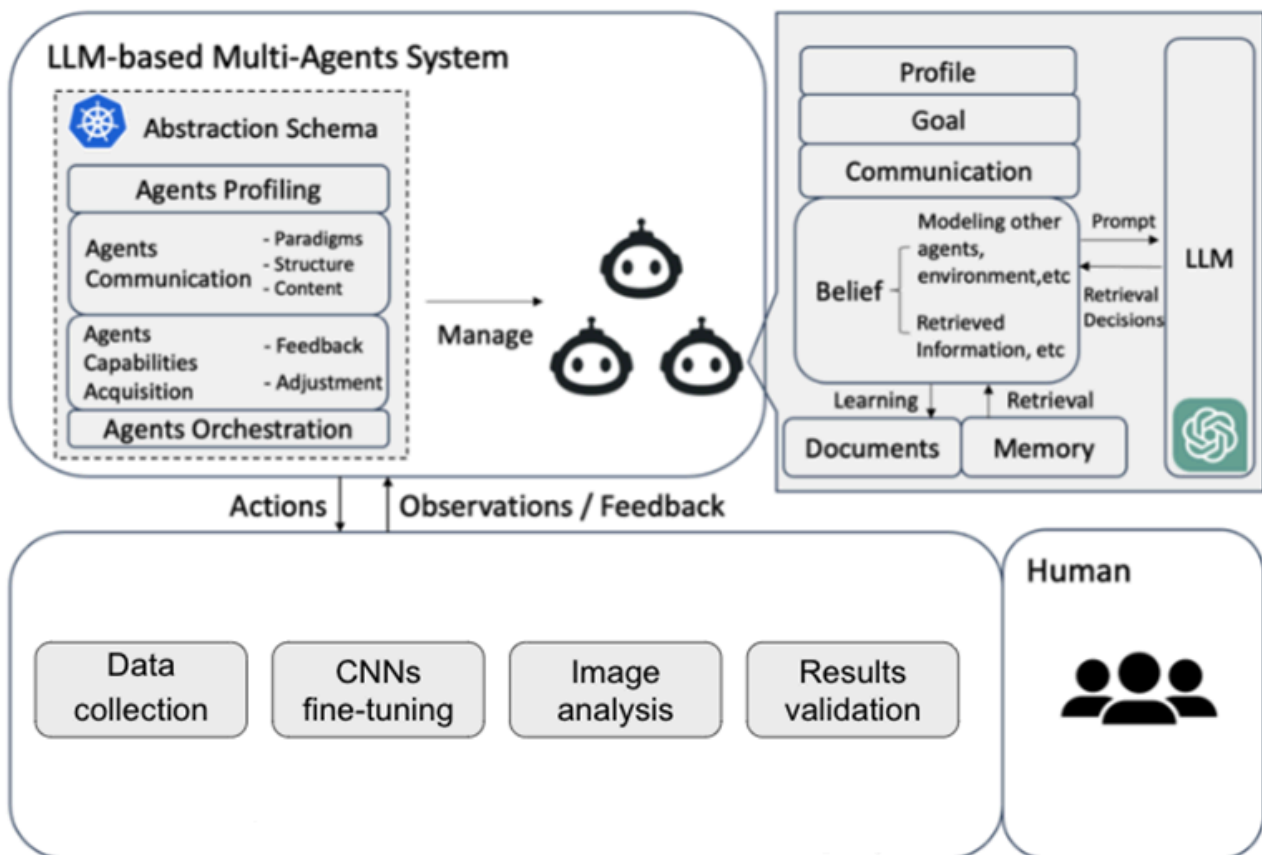
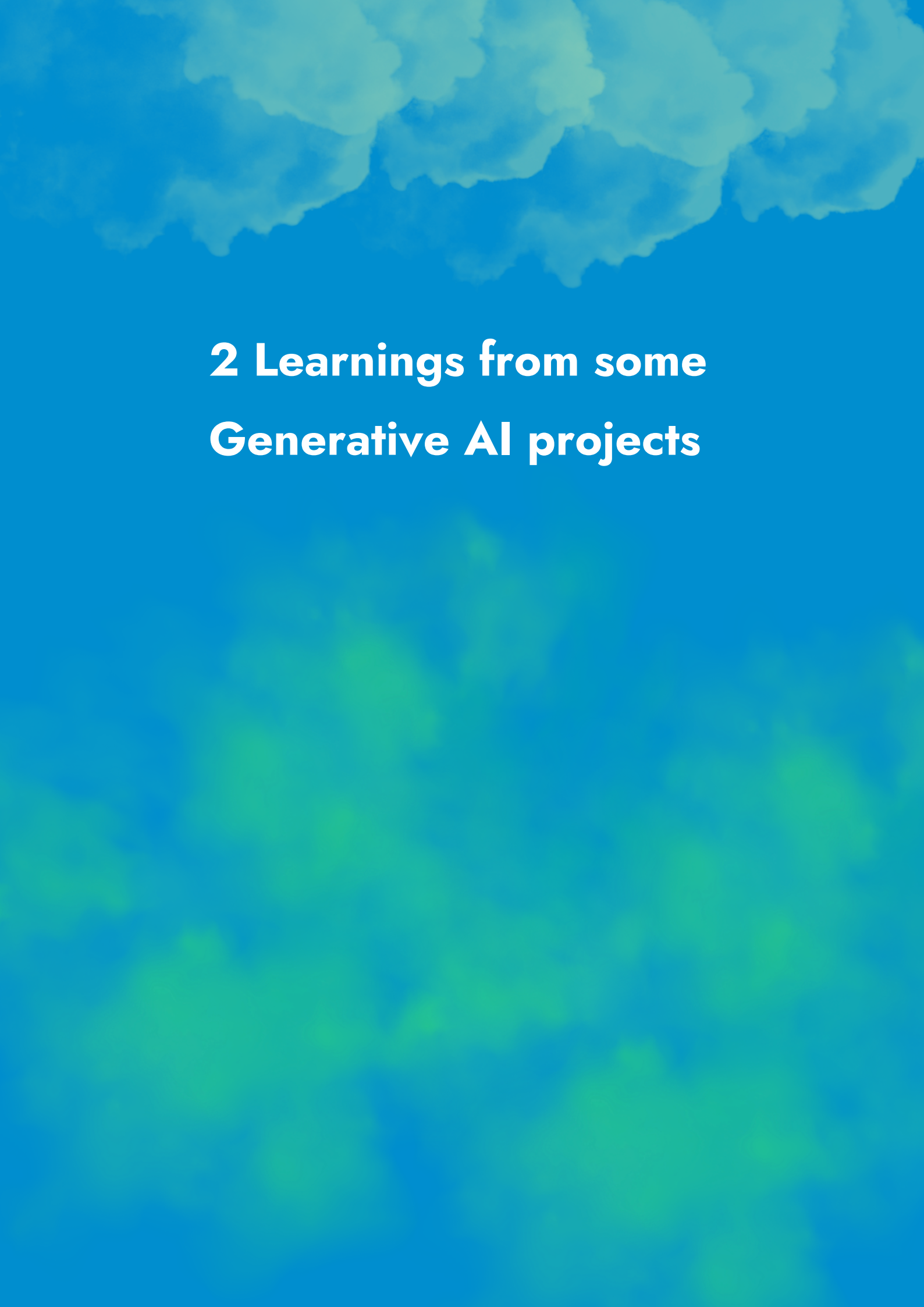


Figure 3 LLMs and multiagent systems in support of automation of workflows

LLM-based Multi-Agents can be proposed to leverage the collective intelligence and specialised profiles and skills of multiple agents. Compared to systems using a single LLM-powered agent, multi-agent systems offer advanced capabilities by: i) Specialising LLMs into various distinct agents, each with different capabilities; ii) enabling interactions among these diverse agents to simulate complex real-world environments effectively.

Generative and Predictive AI can, thus, potentially be used for challenging applications such as the tracking of Tropical Cyclones. The establishment of robust pipelines and workflows can be fed from data spaces with high quality, authoritative and harmonised data sets supporting the collection, generation and validation of such impactful and life-threatening decision-making insights. The pipelines involve complex systems, models and data/compute infrastructures. There should be a collective responsibility from all actors involved on such pipelines to optimise and evolve the systems to deliver the best possible outcomes.

The background of the slide is a bright blue sky with soft, white, fluffy clouds. The clouds are more concentrated at the top and bottom edges, leaving a clear blue area in the center where the text is located.

2 Learnings from some Generative AI projects

2.1 Language Data: OpenGPT-X case

Traditionally, GenAI has focused on creating Large Language Models based on textual data, more specifically, many models were primarily based on English data. **OpenGPT-X** is a BMWK-funded lighthouse project aiming to create large language models covering all of the official EU languages. The project consortium includes research organisations, infrastructure providers, and industry partners. As the project trains large language models from scratch, it faces challenges across all parts of the applications lifecycle. In the following, we map the challenges we faced and lessons learnt in the project so far against the AI model lifecycle:

Legal/ethical requirements and compliance: Properties of the used datasets (such as copyrighted information within web-scraped data or the occurrence of personally identifiable information) pose challenges for trained models to be compliant with current laws and regulations. Different formats and representations of data and metadata (e.g. metadata fields, formats, compression) limit the interoperability between components and restrict re-use of data.

Data spaces could help in normalising the metadata required to identify problematic content by providing clear and complete information for each dataset, e.g. including licence information and dependencies between datasets. Thus, data spaces can prepare the way towards higher standardisation in data governance.

Use case/application requirements from the model lifecycle:

- *Scope:* A clear definition of the expected properties of data and resulting model is essential. This includes the number of model parameters, the required number of tokens to train a model, content languages and the use cases or tasks to be supported. Web pages from CommonCrawl still comprise the majority of the training data, covering these requirements to a large extent. The download, preprocessing, and storage of these datasets requires extensive compute resources. To add diversity to the training data, curated content covering multiple domains, genres, writing styles complement the web data.

Data spaces can help identify relevant compute resources, processing services, and datasets based on clear dataset descriptions and support a consistent processing of datasets with interoperable services.

- *Select/Pre-train:* From the range of potentially useful datasets for pre-training, a subset representing different domains, writing styles and types of information is selected. Training LLMs requires an enormous amount of diverse, high-quality multilingual training data. To ensure high quality, training data is typically filtered according to quality criteria (such as the content length or confidence score for the detected language, or its resemblance to other high quality content). In addition, toxic or offensive content and near-duplicate content is removed. In OpenGPT-X, the preprocessed datasets are sampled (by domain and language) to obtain a more balanced dataset which is less biased towards English.

Data spaces help to identify relevant data and services required for processing the data.

- *Adapt:* Pretrained models are typically fine-tuned for specific domains, tasks, or languages. Fine-tuning a model for the 24 official EU languages poses a challenge as there are few (if any) tuning datasets covering them all.

Data spaces could help identify datasets for specific intended use and could also incorporate services to generate synthetic data on-the-fly. They also provide the technical core to create processes alleviating the lack of scalable multilingual data, e.g. including services automatically translating and manually curating existing tuning datasets.

- *Evaluate:* As the OpenGPT-X models were trained on data in the 24 official EU languages, a corresponding evaluation in 24 languages across a variety of tasks proved difficult, as there are few evaluation datasets covering these languages. As part of the project, new evaluation datasets were created using a semi-automatic approach based on automatic machine translation followed by manual corrections.

Similar to the case of adapting models, data spaces could help in creating processes to generate evaluation data semi-automatically.

- *Deploy:* Within the project, a model playground was created to showcase capabilities of the trained models and facilitate comparison to other state-of-the-art models. Deployed models are typically monitored and might require updates or continuous training, which may increase model performance when new datasets become available.

Data spaces can provide the means to help automate this process.

Data needs: Information about datasets and access to the data are key steps to prepare training data for LLMs. The data processing requires a scalable and robust data pipeline supporting technical infrastructures for different HPCs. Data spaces such as LDS would offer means to identify and access data resources more easily.

Computing needs: Acquiring the data, preprocessing it and training a LLM with the data requires computing resources and access to an underlying technical infrastructure. In the case of training an LLM from scratch, the associated costs (i.e. disk space, CPU and GPU hours) might be prohibitive in most cases. However, data spaces might support a normalised application process for compute resources and thus, a fairer distribution of compute resources.

It should be noted that the storage of metadata (descriptions of data sets), the storage of the physical data close to where it is processed pose as well as the hosting of data space services require resources which need to be allocated for almost every project. Long-term initiatives such as LDS are a step towards data spaces providing a lower entry level in new research projects.

2.2 Public Tenders Data: Tenders Electronic Daily (TED)

Tenders Electronic Daily (TED) is the platform where all public tenders published in EU Member States and European institutions are accessible. With approximately 520,000 public procurement notices published per year that are worth more than €420 billion, TED is a cornerstone of the European public procurement data space. The TED database is available as open data, providing an extremely interesting source for in-depth analysis on public procurements in the EU.

The European Commission Joint Research Centre (JRC) has developed an automated procedure to extract insights from TED for any user-defined topic. The procedure starts from the definition of a first set of keywords performed by the user. Keywords are then enriched using SeTA (Semantic Text Analyser Tool, a text mining algorithm developed at the JRC, trained on hundreds of thousands of EU documents and able to produce a set of "EU-related" keywords for any topic); finally, keywords are validated and finalised by the user.

At this point, the TED database is searched for the list of keywords (which can be combined using boolean operators to optimise the search) to identify the candidate records, i.e. procurement notices that may be relevant to the topic initially defined. Finally, an LLM is used to classify these candidates and determine which of them are indeed relevant. This is done by carefully 'guiding' the LLM through the temperature parameter and the provision of additional context to help maximise the correctness of the classification.

The procedure (using TED for 2021 and 2022) was tested on a set of use cases, for example to identify the presence of procurements relevant to the use of open-source software, open source geospatial software, proprietary geospatial software, open data and open geospatial data, open standards and open geospatial standards.[106]

[106] See Minghini M., Semantic annotation and classification of EU tendering data on open geospatial software, standards and data using Large Language Models. Presentation given at FOSS4G Europe 2024, <https://talks.osgeo.org/foss4g-europe-2024/talk/MTEY TZ/>

The main contribution provided by data spaces - and the public procurement data space in particular - to GenAI, is the (increased) availability, discoverability and accessibility of the TED database. While TED was available under an open access license even before the appearance of data spaces, the actual use of such a database in an efficient and actionable way has been limited. This was mainly caused by the intrinsic complexity of the database and even more by its size, which typically makes it cumbersome to extract relevant insights.

The main users of TED are businesses, which are interested in either finding relevant market opportunities, monitoring procurement trends or simply improving their knowledge of the market. GenAI offers unprecedented opportunities to leverage the TED database to quickly, easily and insightfully extract information on EU public procurements, including in terms of analytics. Typical use cases include answering questions such as how many procurements have been published on a certain subject, when, by whom, in which EU Member State or societal field, whether this has changed over time and how.

In the context of data spaces, processing this database with the help of GenAI can provide great opportunities not only to actors from the private sector, but also to public sector providers of data/procurements, intermediaries and users. It would also contribute to the overall objective of the public procurement data space to improve transparency and accountability of public spending, fighting corruption and improving spending quality. In turn, this opens up opportunities for all other common European data spaces, since European procurements are relevant to all societal domains. Finally, the public procurement data space should also make it easier to link the TED database of public procurements with the related Member States/EU datasets relevant to procurements and the corresponding conditions for access and use, thus offering users an enhanced experience in finding, accessing, consulting and reusing those datasets.



3 Learnings from several sectors based on research activities and projects

3.1 Healthcare domain: Clarify project case

GenAI models, which empower the creation of text and images, are increasingly recognised as a pivotal and transformative tool in the healthcare domain. These advanced technologies promise to revolutionise healthcare by significantly enhancing diagnostic precision, streamlining healthcare delivery processes to reduce costs and time, and ultimately leading to better patient outcomes. Sai et al.[107] presents a comprehensive list of applications of GenAI in the healthcare domain. According to the authors, GenAI has unlocked numerous opportunities in healthcare, with one of its most impactful applications being in medical imaging. Advanced models such as DALL-E are making significant strides in disease detection and diagnosis by analysing patient images, including X-rays, MRIs, and CT scans. Additionally, GenAI is revolutionising the design of molecular structures. It can generate novel compounds, optimise drug candidates, and predict drug properties. Conditional Variational Autoencoders (CVAEs), a specific type of GenAI model, integrate conditional variables with variational autoencoders to learn latent patient information and incorporate conditions pertinent to personalised treatment. GenAI techniques are also transforming medical education through virtual patient simulations, procedural simulations, scenario generation, and haptic feedback.

In clinical trials, GenAI enhances protocol design and validation by simulating virtual trials with various designs, randomisation strategies, or inclusion criteria. It supports stressed individuals by offering personalised treatment plans and therapies tailored to their specific needs. GenAI technologies can simulate and analyse human movement, providing valuable insights into diagnostic, therapeutic, and performance optimisation processes. GenAI has the potential to improve prior authorisation procedures by allowing healthcare practitioners to input patient information into the system.

[107] Sai, S., Gaur, A., Sai, R., Chamola, V., Guizani, M., & Rodrigues, J. J. (2024). Generative ai for transformative healthcare: A comprehensive study of emerging models, applications, case studies and limitations. IEEE Access.

The GenAI models can then quickly analyse this data to prioritise patients based on the severity of their conditions. Furthermore, GenAI models offer deep insights into the complex mechanisms behind adverse drug reactions (ADRs), improving our understanding of drug safety. Lastly, GenAI provides a groundbreaking approach to identifying biomarkers, enhancing clinical judgement, advancing personalised healthcare, and deepening our understanding of diseases.

One notable way in which GenAI is supported in healthcare is through the use of data spaces. Data spaces ensure that data used in AI models adhere to high standards of quality, privacy, and security. They also facilitate the sharing of resources among trusted partners. For instance, in our work involving the classification of clinical notes into categories such as "IgA Nephropathy," "Diabetes," and "Hypertension," data spaces play a crucial role. By enabling the sharing of clinical notes across institutions, data spaces help in creating a rich, diverse dataset that is essential for training robust AI models.

To maximise the benefits of GenAI, data spaces provide standardised, annotated data formats across participating institutions. This standardisation reduces preprocessing costs and ensures uniformity, which is critical for effective model training and reliable outcomes. Additionally, the integration of SemMedDB—a comprehensive PubMed repository of biomedical semantic predications—further enhances the value of data spaces in healthcare. SemMedDB aligns with the concept of data spaces by offering standardised data and structured biomedical knowledge through knowledge graphs and linked data. By linking clinical notes with SemMedDB's curated knowledge, GenAI models can extract and utilise detailed semantic features. This integration leads to more accurate predictions and insights into patient conditions, drug interactions, and potential complications.

In summary, GenAI, supported by data spaces and enhanced by resources like SemMedDB, promises significant improvements in healthcare diagnostics and treatment, offering more precise predictions and better overall patient care.

The work described within the use case was conducted under the **H2020 Clarify project**.^[108] Research on this topic continues by the University Hospital Galway, Ireland and Insight – the SFI Centre for Data Analytics and is funded by the Health Service Executive, Ireland.

3.2 Cybersecurity domain: Addressing Data Availability Issues in Mass Markets

Cybersecurity vendors are increasingly developing AI-based security applications to both enhance the security awareness of users and improve security capabilities. However, one of the main challenges faced by cybersecurity vendors is data availability. For instance, consider the usage of mobile phones. Both private and business customers use mobile phones for confidential or sensitive matters. At the same time, these users handle messages that include links from multiple sources, such as email, SMS, chat apps, and social media, resulting in significant cybersecurity challenges. For a cybersecurity vendor, protecting a private or business customer is a wicked challenge. To analyse the reputation of the links, the vendors would need to evaluate the reputation of the links. However, the messages are encrypted that delimits the vendors ability to protect the customers.

In theory, the concept of data spaces could address these kinds of cybersecurity challenges by creating secure environments where data can be shared and analysed without compromising privacy or security. For instance, we could utilise data spaces for the example of securing mobile users. In this theoretical model, the key parts are:

1. **Secure and Privacy-Preserving Data Sharing Space for Mobile Users:** Establish a data space where mobile users can opt-in to share metadata about the links they receive across various platforms.
2. **Cybersecurity Coalition Data Space:** This part facilitates collective efforts among various cybersecurity vendors to analyse and evaluate link reputations, leveraging shared intelligence within a secure and privacy-compliant environment.
3. **Unified Alert Protocol:** This part of the data space involves a standardised system for communicating potential risks associated with links to mobile users.
4. **Regulatory Compliance:** The part operates within the regulatory frameworks of the EU, ensuring compliance with data protection laws such as GDPR.

However, to implement this theoretical model is not easy in practice. The Part 1, Secure and Privacy-Preserving Data Sharing Space for Mobile Users, is a multifaceted problem and the legal authorities and government will not accept the case. The case would leave too many threats to privacy policies and abuse. On the other hand, these discussions will continue as part of a broader conversation on how to balance user protection with privacy rights in the digital age. This is a call for researchers and authorities to find means to offer cybersecurity vendors a means to protect users and, at the same time, protect mobile users from privacy violations.

3.3 Agriculture domain: SoilWise project

The **SoilWise project**,^[109] develops the **Soil Mission**^[110] data and knowledge repository, which is expected to advance and become part of the **EC JRC European Soil Observatory (EUSO)**^[111]. It also contributes to the implementation of the European Digital Strategy and the European strategy for data and acts as a digital enabler for the European Green Deal.

SoilWise leverages GenAI (GenAI) to support information retrieval, knowledge generation and sharing in soil health monitoring. It is expected that the combined use of Large Language Models (LLMs) and Knowledge Graphs (KGs) in harvesting metadata will allow the SoilWise repository to extract and interpret data from various soil health sources, including all projects under the Horizon Europe **Mission Soil**^[112] initiative, providing advanced catalogue services on data and knowledge products. The combination of Large Language Models (LLMs) and knowledge graphs, as part of GenAI, enhances the capabilities for generating and understanding content and facilitating the creation of a soil-health-knowledge-graph, which captures critical information on soil health indicators, biodiversity, and ecosystem services. For example, SoilWise is using LLMs to enhance metadata with keywords, spatial locations, and translations, improving the organisation and accessibility of soil health data.

The primary stages in the SoilWise GenAI lifecycle are data collection, metadata integration, knowledge graph construction, quality assurance, model training, and optimisation. The challenges in these stages are primarily related to data accessibility, quality, consistency, and semantic interoperability across various data sources. Addressing these challenges is essential to building a reliable knowledge graph that effectively supports soil health monitoring and management.

[109] <https://soilwise-he.eu/>

[110] https://research-and-innovation.ec.europa.eu/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-europe/eu-missions-horizon-europe/soil-deal-europe_en

[111] [https://joint-research-centre.ec.europa.eu/eu-soil-observatory-euso_en#:~:text=The%20EU%20Soil%20Observatory%20\(EUSO\)%20underpins%20EU%20policies%20that%20touch,%2C%20EU%20soil%20protection...&text=The%20EU%20Soil%20Observatory%20\(EUSO\)%20will%20be%20a%20main%20source,reports%20on%20soil%2Drelated%20matters.](https://joint-research-centre.ec.europa.eu/eu-soil-observatory-euso_en#:~:text=The%20EU%20Soil%20Observatory%20(EUSO)%20underpins%20EU%20policies%20that%20touch,%2C%20EU%20soil%20protection...&text=The%20EU%20Soil%20Observatory%20(EUSO)%20will%20be%20a%20main%20source,reports%20on%20soil%2Drelated%20matters.)

[112] https://research-and-innovation.ec.europa.eu/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-europe/eu-missions-horizon-europe/soil-deal-europe_en

Data spaces, such as the Agricultural and Green Deal, can play a critical role, considering that they emphasise **data accessibility, data and knowledge-sharing governance and semantic interoperability**. **By setting up the conditions for accessing persistent, interoperable and high-quality data and knowledge sources, data spaces can ease the harvesting process, enhancing the accuracy and the application of GenAI practices and technologies.**

Conversely, SoilWise can contribute to these data spaces by providing GenAI-based catalogue services, allowing knowledge generation and sharing between various actors in the soil health ecosystem. This capability is crucial for informed decision-making and promotes collaboration, innovation, and sustainable soil management practices.

The background of the page is an abstract, marbled pattern in shades of green and white. The pattern consists of flowing, organic shapes that resemble liquid or smoke, creating a textured and dynamic visual effect. The colors range from light, almost white, to various shades of green, including some darker, more saturated tones.

Annex A - EU policy framework and relevant initiatives

Although GenAI has been around in some form for decades, the newly widespread availability of publicly-accessible and consumer-facing GenAI models has quickly reshaped the digital landscape since they entered the market at scale in late 2022.[113] These powerful GenAI models have captured public attention and attained new levels of performance, demonstrating the potential for various real-world applications and triggering a wave of research and development, as well as large corporate investments. [114] GenAI is likely to have a significant impact across a wide range of industries as it finds its way into products, services and processes, enabling the creation of new business value, efficiency gains and innovations, and offering innovative solutions to shared challenges. However, a technological and investment gap between the EU and other key players in strategic technologies persists and the market for GenAI tends to be concentrated in a few powerful private companies – mostly operating from outside the EU.[115] For this reason, the EU increasingly advocates a message of technological or digital sovereignty.

Therefore, there is a need for enabling digital and industrial policies and a supportive governance framework at EU level, in order to guide and stimulate the development and deployment of GenAI within the Union. The EU has been actively working to establish such policies, as well as a comprehensive set of standards and legislation, also addressing data as a key foundation of AI (including GenAI). The Commission has presented a vision for Europe's digital transformation by 2030[116] and has set ambitious targets aimed at strengthening digital sovereignty through specific actions on data, technology and infrastructures. Various strategic initiatives have been taken, with the intention to set out an enabling framework for the data economy and to boost Europe's technological and industrial capacity in AI and its uptake. These policies reflect the EU's approach to balancing the promotion of innovation and industry with the need to ensure these technologies are developed and utilised in a way that is safe, ethical, and respects fundamental rights. The following sections will outline this enabling framework, along with relevant initiatives (projects and programmes), noting that the policy and regulatory context continues to rapidly evolve.

[113] Future of Privacy Forum (May 2024), Navigating Governance Frameworks for Generative AI Systems in the Asia-Pacific, p. 4.

[114] See Stanford Human-Centered Artificial Intelligence (HAI) (2024). Artificial Intelligence Index Report, https://aiindex.stanford.edu/wp-content/uploads/2024/05/HAI_AI-Index-Report-2024.pdf and World Intellectual Property Organisation (WIPO) (2024), Generative Artificial Intelligence. Patent Landscape Report. Geneva: WIPO. <https://doi.org/10.34667/tind.49740>

[115] European Commission (2024), Science, Research and Innovation Performance of the EU 2024: A competitive Europe for a sustainable future, https://research-and-innovation.ec.europa.eu/document/download/67d10296-11c3-48c9-9c3b-b165a3d946ad_en?filename=ec_rtd_srip-report-2024.pdf

[116] European Commission Communication, 2030 Digital Compass: the European way for the Digital Decade, COM(2021) 118 final.

A.1 The EU vision for a data-agile economy

Data are fundamental resources for the development and implementation of AI (including GenAI) models and systems - access to high quality and diverse datasets is of critical importance for model generation, training and improvement. Data governance frameworks thus play a crucial role in the exercise of digital sovereignty. The EU has established a robust framework of data-related policies and standards aimed at promoting wider availability, sharing and reuse of data, while at the same time protecting personal data, ensuring data security, and promoting the ethical use of data. These policies and standards are particularly relevant for the development and deployment of GenAI, which often relies on large datasets for its training and operation, and which must comply with EU data policies to ensure lawful data processing.

The high standard of data protection promulgated in the EU's General Data Protection Regulation (GDPR 2016/679), is coupled with policies that aim to develop a common European data market to spur innovation, foster economic growth and advance societal development. The **European strategy for data**, released in 2020, sets a goal for the EU to become a leading role model for a society empowered by data to make better decisions, both in business and the public sector.[117] By ensuring greater data availability and fostering cross-sector collaboration, the strategy can help significantly boost the capabilities of GenAI systems that rely on such data. Announcing the political priorities for the 2024-2029 European Commission, President von der Leyen emphasised that "To support the development of AI and other frontier technologies, Europe needs to exploit the untapped potential of data [...] Europe needs a data revolution.".[118]

To implement this vision, several cross-sectoral legislative instruments have been adopted, with the aim of ensuring fairness in the digital environment, stimulating a competitive data market, opening up opportunities for data-driven innovation and making data more accessible for all.

[117] European Commission (2020), Communication on A European strategy for data, COM(2020) 66 final

[118] Ursula von der Leyen (July 2024), 'Europe's Choice, Political priorities for the 2024-2029 European Commission', https://commission.europa.eu/document/download/e6cd4328-673c-4e7a-8683-f63ffb2cf648_en?filename=Political%20Guidelines%202024-2029_EN.pdf

These include, inter alia, the **Data Governance Act**,^[119] the **Data Act**,^[120] and the **Implementing Act on High Value Datasets**.^[121] Specifically relating to public sector data, the **Interoperable Europe Act** also promotes data sharing, reuse and collaboration. ^[121] (more on the above can be found on chapter 3 of Part 1).

These complementary pieces of legislation serve to promote a competitive and trustworthy data economy, enabling AI developers to access diverse datasets while respecting data ownership rights. GenAI models and systems stand to benefit from this access to a wider range of high-quality datasets, including both personal and non-personal data, which will enhance model training and innovation, while ensuring that data sharing practices are compliant with EU regulations. While the Data Act does not explicitly mention AI (though it does reference machine learning), it may directly impact GenAI companies and providers, as many connected devices rely on the application of AI technologies.

In this context and complementing the horizontal actions towards a single market for data, while also taking note of technical advancements, **common European data spaces** emerge as key instruments to foster the European data economy. The European strategy for data announced the development of data spaces in strategic economic sectors and domains of public interest (e.g., health, agriculture, Green Deal, energy, mobility, manufacturing, tourism, etc.) These were subsequently supplemented by several additional sector-specific data spaces, and are aimed at creating large, interoperable data environments across sectors to enable data sharing and collaboration (which can also be leveraged for GenAI training and innovation).^[122] **The 2023 Annual Single Market report** highlights the ambition to create a data-driven Single Market where interoperability within and across data spaces is ensured^[123] and the **Digital Europe Programme (DEP)** will provide the necessary support for the development of these data spaces.

All these data policy initiatives represent key pillars to support the EU's digital transformation and its pursuit of the digital targets for 2030. They promote innovation by providing access to high-quality, ethically sourced data and will collectively help ensure that GenAI technologies develop, and are used, in a manner that respects data privacy, fosters innovation, and upholds ethical standards.

[119] Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724.

[120] Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828.

[121] Regulation (EU) 2024/903 laying down measures for a high level of public sector interoperability across the Union (Interoperable Europe Act).

[122] European Commission Staff Working Document on Common European Data Spaces, SWD(2022) 45 final.

[123] European Commission, Commission Staff Working Document 2023 Annual Single Market Report: Single Market at 30, SWD(2023) 26 final.

A.2 EU AI-related policy landscape

The EU has actively been setting up a range of policy and regulatory measures, as well as standards, to guide and support the development, deployment and use of AI. For these purposes, GenAI is considered to fall under the broader umbrella of AI, although it also has certain specificities. The EU aims to foster an innovative European AI ecosystem where startups and innovators can work closely with industrial users, attract investments and have access to the key ingredients of AI - data, computing, algorithms and talent. In addition, the EU aims to boost innovative applications for AI in Europe's industrial ecosystems, while upholding European values, tackling risks and promoting the responsible use of AI (including GenAI). The focus is on Trustworthy AI, defined as AI that is legally compliant, ethically adherent, and socio-technically robust throughout the system's entire lifecycle. [124]

As pointed out in the recent 'Draghi' report on the future of European competitiveness, AI – and particularly GenAI – is an evolving technology in which EU companies still have an opportunity to carve out a leading position in selected segments.[125] According to the 2024 Commission Communication on boosting startups and innovation in trustworthy AI,[126] mastery of the latest developments in GenAI will become a key lever of Europe's competitiveness and technological sovereignty. A report by the Centre for European Policy Studies (CEPS) points to the need to promote European technological sovereignty by developing trustworthy AI models that are compliant with the principles and values rooted in the EU treaties, ethically aligned and socio-technically robust.[127] It cites research estimating that 73% of cutting-edge LLMs are currently being developed in the US and another 15 % in China.[128]

[124] European Commission (2019), Ethics Guidelines for Trustworthy AI.

[125] European Commission (2024), 'The Future of European Competitiveness – A competitiveness strategy for Europe', September 2024.

[126] European Commission, Communication on boosting startups and innovation in trustworthy artificial intelligence, COM 2024 (28) final.

[127] CEPs (2024), Towards a European large-scale initiative on Artificial Intelligence: What are the options?, July 2024 - 11.

[128] Council of the EU (24 April 2023), ChatGPT in the Public Sector –overhyped or overlooked?, research paper by the Analysis and Research team (ART).

During the past few years, the Commission has been facilitating and enhancing cooperation on AI across the EU to boost its competitiveness and ensure the trustworthy development and use of AI based on EU values. In 2018, it published a **Coordinated Plan on the Development and Use of Artificial Intelligence**, designed to accelerate investments, define strategies and align AI policies among its member states.[129] An updated Plan was released in 2021, putting forward a concrete set of joint actions on how to create EU global leadership on trustworthy AI.[130] Also in 2021, **the White Paper on Artificial Intelligence**[131] presented policy options to ensure that the development of AI is trustworthy, secure, and in line with the values and rights of EU citizens.[132] It introduced the concepts of “ecosystem of excellence” along the entire value chain of AI adoption, and “ecosystem of trust” to give citizens, businesses, and public organisations the highest possible confidence in using AI.

The first comprehensive EU regulatory framework laying down harmonised rules on AI, the **Artificial Intelligence Act**, entered into force on 1 August 2024[133] This new regulation emphasises an approach that is risk-based and shaped by EU values, ensuring both safety and fundamental rights protection. It aims to classify AI systems based on the risk they pose (e.g., unacceptable, high, limited, and minimal risk). Within the Act, special attention is given to providers of high-risk AI systems and of General-purpose AI models, particularly those posing systemic risks. This is underpinned by the powers given to the Commission under the AI Act, including the ability to conduct evaluations of GPAI models, request information and measures from model providers, and apply sanctions. There are also proposals for an **AI Liability Directive** and an **AI Product Liability Directive**.

The **Digital Services Act (DSA)**[134] and the **Digital Markets Act (DMA)**[135], which entered into force in 2022, are designed to create a safer and more open digital space by regulating digital platforms and ensuring fair competition. These acts could affect GenAI tools integrated into digital platforms, particularly in terms of content moderation, transparency of algorithms, and competition. For instance, AI-generated content on large platforms must comply with content moderation rules under the DSA.

[129] Coordinated Plan on the Development and Use of Artificial Intelligence (COM (2018) 795 final).

[130] Coordinated Plan on Artificial Intelligence 2021 Review, COM(2021) 205 final.

[131] COM(2020) 65 final.

[132] White Paper on Artificial Intelligence – A European approach to excellence and trust (COM/2020/65).

[133] Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).

[134] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act).

[135] Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act).

The updated **EU Industrial Strategy** indicates the need for investment and innovation to deliver innovative technologies and illustrates significant business opportunities in these areas.

Enabling measures for AI at EU level also include substantial financial stimulus. **Horizon Europe** and the **Digital Europe Programme** provide significant funding opportunities for research and innovation in AI, including GenAI technologies. Furthermore, through the '**AI Innovation Package**' launched in January 2024, the Commission set out a strategic investment framework and launched a package of measures to support European startups and SMEs in the development of trustworthy AI that respects EU values and rules. The **GenAI4EU initiative** is part of this package and will facilitate the development of novel use cases and emerging applications in 14 industrial ecosystems in Europe. Application areas include robotics, health, biotech, manufacturing, mobility, climate and virtual worlds.

The accompanying **Communication on boosting start-ups and innovation in trustworthy AI**[136] lays out a series of actions and investments that will help startups and industries in Europe fulfil their potential of becoming global frontrunners in trustworthy advanced AI models, systems and applications. It also promotes acceleration of the development and deployment of European data spaces, accessible to the AI community, recognising data as a crucial resource for training and improving AI models. This AI Innovation package will generate an additional overall public and private investment of around €4 billion until 2027.

The development of capabilities regarding GenAI also depends on physical infrastructures capable of supporting the generation and processing of large volumes of data. As part of the AI Innovation package, an amendment to the **EuroHPC Regulation**[137] to set up AI Factories entered into force in July 2024, enabling the creation of a new pillar of activities for the **EuroHPC Joint Undertaking** and representing a significant step forward in developing **GenAI4EU** capacities and resources. Expected to be deployed in 2025, the AI Factories will consist of dedicated AI-supercomputers to facilitate fast machine learning, in particular for training of GenAI models. They will offer a one-stop shop for startups and innovators and enable the development of powerful models to underpin a variety of emerging AI applications, as well as facilitating the development of smaller, more efficient models, compared to existing LLMs.

[136] COM 2024 (28) final.

[137] Council Regulation (EU) 2021/1173 of 13 July 2021.

The **European Digital Infrastructure Consortium (EDIC)** is a new legal instrument made available to Member states within the framework of the **Digital Decade Policy Programme 2030** to speed up and simplify the setup and implementation of **multi-country projects** to install shared digital infrastructures in priority areas. These transformative joint projects will be led and funded by Member States, with the support of the EU. One of the first projects relates to an **Alliance for Language Technologies (ALT-EDIC)**,^[138] the mission of which is to develop a common European infrastructure in Language Technologies, particularly focusing on LLMs and embracing the linguistic diversity of the Union. By pooling resources, the members should achieve the critical mass of data and other resources needed to create and finetune LLMs, improve supply of high-quality language data, contribute to improving the capabilities of AI models in addressing the linguistic needs of the smaller Member States and language communities with limited language data.

Another relevant EDIC project, **Networked Local Digital Twins towards the CitiVERSE (LDT CitiVERSE EDIC)**, will connect existing local digital twins created in member cities, forming the basis of the EU CitiVERSE.^[139] The CitiVERSE will focus on advancing GenAI-based applications for virtual reality in smart cities, compatible with EU laws and values.

As a further initiative to strengthen the EU's technological infrastructure, the **European Chips Act**^[140] is a legislative framework aimed at bolstering Europe's competitiveness and resilience in semiconductor technologies and production. While the act itself is specifically targeted at the semiconductor industry, its impact is likely to create a more favourable environment for the research, development and deployment of advanced AI applications, including GenAI. Advancements in R&D in the semiconductor sector could lead to more efficient and powerful chips, which in turn would benefit AI applications by allowing for faster processing and more sophisticated model creation: Gen AI could see significant performance improvements as a result. The Chips Act could also help ensure the hardware used in data centres where GenAI models are trained and deployed is reliable and not subject to external supply chain vulnerabilities. Furthermore, the Act aims to foster ecosystems that enable innovation, including partnerships between industry, academia, and research institutions. However, bearing in mind the sustainability challenges mentioned in Section 2, it will be important to integrate energy policy with the Chips Act to mitigate the environmental impact of increased chip production (e.g., by promoting the use of renewable energy sources for manufacturing).

[138] [Implementing Decision \(EU\) 2024/458](https://eur-lex.europa.eu/eli/dec_impl/2024/458/oj), https://eur-lex.europa.eu/eli/dec_impl/2024/458/oj

[139] Implementing decision (EU) 2024/459, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202400459

[140] Regulation (EU) 2023/1781 of the European Parliament and of the Council of 13 September 2023 establishing a framework of measures for strengthening Europe's semiconductor ecosystem and amending Regulation (EU) 2021/694 (Chips Act).

Through all these policies, actions and legislative measures connected to the digital transformation, the EU is seeking to promote investment and innovation in the crucial domain of AI (and GenAI) to strengthen its economic competitiveness and strategic autonomy, while at the same time aiming to improve quality of life, resilience, sustainability of industry and enable the green transition. These policies and legislation also influence European standardisation efforts in the field of AI and data, as we will see in the following section. Together they form a comprehensive landscape of laws, policies and standards designed to foster innovation, while protecting the rights and safety of EU citizens in the digital domain. As the technologies continue to evolve, the EU is likely to update and refine its policies, standards and legislation to ensure they remain aligned with its values and objectives and are fit to address new challenges.

A.3 The role of harmonised standards in policy implementation

There is a strong connection between EU policies and standards, where policy makers are working together with standards bodies, industry, and other relevant stakeholders to prepare the necessary technical environment for implementation of policy and legislative frameworks. For certain policies and regulations, the EU may request the creation of "harmonised standards." [141] In the context of data and AI, the EU has been actively working to translate policies and legislation into standards to ensure the technologies developed and deployed within the EU comply with its values and legal requirements.

European standards, while voluntary in nature, are important instruments to advance technical harmonisation in the areas of AI and data, and to ensure a high level of protection of safety and fundamental rights for EU citizens throughout the Union. Standards can also support the establishment of equal conditions of competition and a level playing field for the design and development of AI (and GenAI) systems. When products, services, or systems comply with European standards, they are presumed to be in conformity with the corresponding EU policies and legislation.

Recognised European Standardisation Organisations (ESOs), such as the European Committee for Standardization (CEN), the European Committee for Electrotechnical Standardization (CENELEC), and the European Telecommunications Standards Institute (ETSI) play an important role in developing technical standards, based on standardisation requests received from the European Commission. As part of the **EU strategy on standardisation**, a **High-Level Forum on European Standardisation** [142] was established as a multi-stakeholder platform in January 2023 and is committed to identifying standardisation priorities and aligning standardisation actions with European policy priorities, industrial innovation and investment activities.

[141] A harmonised standard is a European standard that is developed by a recognised European Standards Organisation and that has a legal basis in EU legislation.

[142] <https://single-market-economy.ec.europa.eu/single-market/european-standards/standardisation-policy/high-level-forum-european-standardisation>

These standardisation bodies are developing detailed technical specifications and guidelines that address various aspects of data and AI, such as data interoperability, data governance and data quality, as well as AI system robustness, transparency and risk management. For example, in connection with the AI Act, the European Commission has made a standardisation request to the ESOs in support of Union policy on AI.[143] This request concerns standards to advance technical harmonisation in the field of trustworthy AI, especially as regards the requirements for high-risk AI systems laid out in the AI Act. The **CEN-CENELEC Joint Technical Committee (JTC) 21 on Artificial Intelligence**[144] is currently developing these standardisation deliverables which, in the future, would be able to provide manufacturers with the presumption of conformity with the AI Act. In addition, there are already some existing European standards, including sectoral standards, that address various aspects relating to AI and data.[145]

In relation to the EU Data Act, a draft standardisation request as regards a European Trusted Data Framework is in preparation to be issued to the ESOs for harmonised standards satisfying the essential requirements laid down in Article 33(1) of that Act. The EU Standardisation Work Programme 2025 also refers to this Trusted Data Framework to ensure trusted, legally compliant data use and sharing across parties, including data intermediaries and altruism organisations. This encompasses data catalogue standards for publication and discovery of data assets, ontology standards for integration and use of shared data, and common KPIs to guide European data spaces towards cross-domain interoperability, enhancing transparency and usage. All actions will be coordinated closely with the European Data Innovation Board (EDIB), which can also advise on the prioritisation of cross-sector standards,[146] and, where relevant, the Data Spaces Support Centre.

It is important to keep in mind that the standardisation landscape is dynamic, and new requests can be made as the regulatory environment evolves. Adherence to agreed EU standards for data and AI is also expected to be a benefit of European data spaces, for example in distributed value chains. By providing a common framework for data sharing, data spaces can help harmonise technical and legal standards across actors and sectors and thereby facilitate access to data from different sources. Model providers also need inputs from multiple perspectives and standards can help with this, due to the collaborative nature of their development and access to inputs from a diverse community.

[143] Commission implementing decision of 22.5.2023 on a standardisation request to the European Committee for Standardisation and the European Committee for Electrotechnical Standardisation in support of Union policy on artificial intelligence, C(2023)3215 – Standardisation request M/593.

[144]https://standards.cencenelec.eu/dyn/www/f?p=205:22:0:::FSP_ORG_ID,FSP_LANG_ID:2916257,25&cs=1827B89DA69577BF3631EE2B6070F207D

[145] Soler Garrido, J., Fano Yela, D., Panigutti, C., Junklewitz, H., Hamon, R., Evas, T., André, A. and Scalzo, S., Analysis of the preliminary AI standardisation work plan in support of the AI Act, EUR 31518 EN, Publications Office of the European Union, Luxembourg, 2023,

[146] For a description of the role of the EDIB, see Article 30(h), of Regulation (EU) 2022/868.

A.4 Relevant European actions (projects & programmes)

A.4.1 HPC/Computing community

GenAI requires high-performance computing (HPC) capabilities to train and run complex models. Other methods spanning the compute continuum, such as cloud and edge services are also fundamental enablers.

At European level, many large AI models are being developed using jointly and publicly owned computing infrastructures via the EuroHPC Joint Undertaking (JU) initiative. Launched in 2018, this is a joint initiative between the EU, European countries and private partners to develop a World Class Supercomputing Ecosystem in Europe. It allows the EU and the EuroHPC JU participating countries to coordinate efforts and pool their resources to make Europe a world leader in supercomputing. This European effort represents a competitive advantage for GenAI-related developments as HPC provides the computational power essential for groundbreaking AI research and applications. Many European supercomputers such as LUMI in Finland,[147],are already very suitable for AI: about 50% of LUMI's capacity is now directed towards AI, to train deep neural networks for different purposes - especially for LLMs - but also for things like image recognition.

As achieving excellence in GenAI requires a comprehensive ecosystem, more than just the hardware, it is important to integrate the infrastructure with skilled expertise and strong collaboration with different national and international partners, from startups to data providers and the scientific community. In this direction, the EuroHPC JU launched in 2024 an "Access Call for AI and Data-Intensive Applications", designed to serve industry organisations, small to medium enterprises (SMEs), startups, as well as public sector entities that require access to supercomputing resources to perform AI-related and data intensive activities.[148]

As mentioned in A.1.2, the EuroHPC JU has also been working this year on a new regulatory effort towards widening AI development capabilities with the introduction of the "AI factories" Amendment, allowing it to operate and acquire AI-optimised supercomputers by upgrading existing systems and procuring new machines.

Also of relevance in this context is **HiPEAC (High Performance, Edge And Cloud computing)**[149], an initiative to bridge industry and academia in computing systems. It provides a hub and network for specialised European researchers and industry representatives in computing systems (currently there are around 2,000 members).

[147] <https://www.lumi-supercomputer.eu>

[148] https://eurohpc-ju.europa.eu/eurohpc-ju-access-call-ai-and-data-intensive-applications_en

[149] <https://www.hipeac.net/>

A.4.2 AI on Demand Platform (AloDP)

Launched in early 2024, the primary goal of the **DeployAI**^[150] European flagship project, which includes multiple participants from across 13 EU countries, is to build, deploy, and launch a fully operational AI-On-Demand platform (AloDP) promoting trustworthy, ethical, and transparent European AI solutions for a seamless implementation in industry and the public sector. It aims to leverage and boost the European AI research landscape, and to serve as a bridge and catalyst between the European AI research and industry, facilitating collaboration and knowledge exchange.

The AloD Platform is vendor independent and is designed to ensure a fair and democratic access to cutting-edge AI technologies made in Europe, including GenAI and LLMs, enabling innovation at a large scale. The vision is for it to become an AI “Lighthouse Project”, on shaping a standard AI infrastructure for all stakeholders in the European industrial ecosystem. Following and emphasising European values, such as openness, transparency, and trustworthiness, it is intended to serve as a one-stop shop for readily usable, responsible and easy to integrate AI tools, and as an enabler and motivator for the European AI Innovation Ecosystem. It will play a prominent role in supporting the AI Factories and the GenAI4EU initiative, as part of the AI Innovation Package, and will support European enterprises (start-ups, SMEs, large industry) in using advanced European AI technologies and ensuring AI sovereignty, especially in relation to GenAI. It will also foster collaboration with the European Digital Innovation Hubs (EDIHs), Testing and Experimentation Facilities for AI (TEFs), Data Spaces, EuroHPC centres, and HPC/Cloud/Edge infrastructures and standardisation organisations.

Fundamentally, the project has two main goals:

1. Firstly, to create a European platform for dissemination and use of European AI technologies, with a focus on GenAI models. To achieve this, the following aspects will be addressed:
 - 1.1. creation of a marketplace/infrastructure for European AI technologies for all market participants, with a particular focus on SMEs, AI start-ups and public clients.
 - 1.2. integration of the Gaia-X trust framework and compliance rules; collaboration with data spaces and EuroHPC Centers.
 - 1.3. promotion of AI trustworthiness and EU regulatory compliance, and
 - 1.4. provision of AI development environments based on European technologies.
2. Secondly, the establishment of a European entity to operate and expand the AloDP.

[150] <https://deployaiproject.eu/>

A.4.3 AI Testing and Experimentation Facilities

In the same way as common European data spaces are poised to become an important pillar in unlocking innovation and adoption for AI-driven applications, powered by datasets that are widely available, interoperable and rich in content and quality - a similar strategy for supporting AI-driven applications is shared by **AI Testing and Experimentation Facilities (TEFs)**,^[151] but with a focus on the validation of tested solutions before they make it to market.

Both data spaces and TEFs feature as important elements of the European strategy for data and will work in synergy to facilitate an easier path to market and wider adoption of AI in strategic sectors such as health, manufacturing, agrifood and smart cities.

The TEFs deploy, exploit and make available a wide network of both physical and virtual facilities. They produce and have access to validation data from the many testing and experimentation services delivered to their customers' near-commercial solutions. These data will be collected in real life conditions, according to rules and procedures specific to each TEF.

Concerning data spaces, having the means to set access and usage policies is an important aspect, further enriched by compute-to-data opportunities, where execution environments are created close to where the data are stored, making it possible to run algorithms and solutions without "opening" proprietary datasets. Thus relying on a data space infrastructure can help secure data sovereignty and pave the way towards valorisation and monetisation of data assets, contributing also to the future business sustainability of TEFs.

TEFs are established to foster AI innovation through services that validate and demonstrate specific aspects of a proposed solution. Expanding the horizon further, the prospect is opening up towards leveraging a combination of facilities, data assets and technologies for deploying GenAI techniques, extracting cause-effect relationships from experimentation results and from sectorial data spaces in a much more effective way, thereby providing reusable models that can make subsequent experimentations easier and shorter.

While "traditional" LLMs might experience only limited benefits from TEFs - mostly in the text-to-text generators that materialise advanced user-to-application interactions, or in the interpretation of scientific outcomes - substantial innovation opportunities are expected for GenAI models, taking advantage of the extensive multimodal datasets progressively produced by TEFs and that are integrated in sectoral data-sovereign data spaces. Pushing those ideas further, should "in-field" testing and experimentation activities struggle to produce or have access to rich enough datasets, GenAI could enrich TEF (and data spaces) catalogues, e.g., through the creation and exploitation of synthetic datasets.

A.4.4 AI Factories

Linked to the EuroHPC new regulation and the AI innovation package described in section 3.1, AI Factories are entities which provide an AI supercomputing service infrastructure and will build open AI ecosystems formed around EuroHPC supercomputing facilities. It is expected that a number of AI Factories will be established across Member States to serve European and national AI communities. AI Factories provide what is needed for development of European general purpose AI models and other emerging AI applications or data driven applications.

Every AI Factory needs to be linked with the national AI strategies and data policies, as well as to the selected AI ecosystems. AI Factories are centred around European AI supercomputers, but the AI Factory concept is much wider. The three key elements are compute, talent and data. Due to the current political and stakeholder focus on the AI Factories, a specific case study on this topic is included in Section 5 of this paper.

A.4.5 Other relevant initiatives in the larger digital ecosystem

Finally, there are several complementary efforts in the broader digital ecosystem that, either directly or indirectly link to data space developments, but which are also relevant for GenAI. This includes, inter alia: Open Data initiatives (e.g., the Open Data Portal, Open Data for AI projects of DEP); Destination Earth Initiative; the European Digital Infrastructure Consortia (EDICs) - see A.1.2; the European Open Science Cloud (EOSC); and Important Projects of Common European Interest (IPCEIs) - Next Generation Cloud Infrastructure and Services (CIS). However, this list is indicative only, and not exhaustive.



Annex B - Technological and industrial trends

This Annex provides some information on technological and industrial trends influencing the development, deployment and adoption of GenAI, particularly within the EU context.

B.1 Technology trends and enablers for GenAI

Since late 2022, there have been significant advancements in GenAI and to some extent, these fast-paced developments have taken many by surprise. Progress has been spurred by recent innovative research and technical advancements in the field, some examples of which are outlined below. Moreover, the amount of data available that can be used to train GenAI models is rapidly expanding (also making data governance, standardisation and regulation crucial to ensure data quality).

The impressive growth and the rapid pace of GenAI development and adoption is largely due to (i) the availability of extensive computational infrastructures, (ii) innovative training algorithms, and (iii) access to large amounts of data. For instance, Large Language Models (LLMs) like Open AI's ChatGPT and Mistral AI's models are trained on a wide array of text, often derived from the Internet.

LLMs and knowledge graphs: 2023 saw a large explosion in the production of LLMs and Multimodal Large Language AI Models (MM-LLMs) due to the introduction of ChatGPT during the previous year. This trend is expected to continue as GenAI increasingly becomes mainstream, and a growth in the production of LLM-based outputs (e.g., videos, images) is also expected. LLMs have already been integrated into multiple sectors, ranging from finance to journalism, health and beyond - they are especially useful for large use cases that require a lot of data.

With this expansion of the GenAI field and the increased need for training data comes the importance of representing and organising knowledge in a smart manner, e.g. using Knowledge graphs (or semantic networks) consisting of nodes, edges and labels.[152] There have been important improvements in knowledge graphs since the introduction of the Semantic Web and they can be one of the enablers of trustworthy GenAI, allowing the mapping of information and data provenance. They have already become integral to AI-based search systems, and in industries such as retail, healthcare, finance and entertainment. Moreover, AI-based solutions such as information retrieval tools, recommender systems and question-answering systems largely benefit from knowledge graphs.

Small models: As the training and development of LLMs is data intensive and costly, developers and companies are increasingly shifting from large models to smaller ones that require less resources. These smaller models require less computing power and are therefore deemed to be more cost efficient and environmentally sustainable. The benefit of these models for companies lies in their ability to perform narrower, more task-specific functions. There are already a handful of use cases in this direction, for example in the customer service field.

Multimodal AI: Recently the focus has shifted from LLMs to other data types or modes - primarily image, video, speech and audio. A simple example of multimodal AI would be a model that is provided with an image as input and produces a textual description of what can be seen in that image. Language models still play a central role as, for example, image or video generation is typically controlled with text-based prompts. The cutting edge of development is currently in video generation models, where consistency and length of the output has proved a challenge in the past. The latest models such as OpenAI's Sora have taken great strides in solving these issues. Meanwhile, the quality of image generation has continued to improve and some teething issues, such as people having the wrong number of fingers, have largely been overcome. One undesirable outcome of these recent developments is the emergence of so-called 'deep fakes', which refer to AI generated media - such as images, videos or speech - that are intended to mislead the recipient into believing those media represent real life situations.

B.2 Industry trends for GenAI

This section considers the practices, attitudes and dispositions of enterprise leaders towards GenAI adoption, addressing first the issue of resources, then turning to models (i.e., closed versus open source) and finishing with the type of use/business case addressed and overall adoption levels. It draws on numerous surveys conducted in relation to this topic, and posits that, in addition to traditional broad-based data and statistics on adoption levels, having information about the “gut feelings” of industry leaders is also of relevance, as they can be important contributors to shaping the landscape of adoption.[153] The analysis is subsequently expanded to assess how different industrial sectors are expected to benefit from GenAI adoption, zooming in especially on the European landscape.[154]

The overall picture is seen to be largely positive. In summary, GenAI is becoming a “more familiar” ingredient of enterprise life with business leaders willing to increase spending on it. In addition, corporates seem to have clear ideas about their priorities concerning use cases (building internal applications, as of now), as well as about the closed- versus open-source divide (increasing interest for the latter) and the level of impact they believe GenAI can bring to their business (high, even though precise metrics are still missing).

Resources needed for GenAI: According to a survey by American venture capital firm Andreessen and Horowitz, the budget for GenAI seems to be growing from an average of 7M\$ in 2023 to an expected 18M\$ in 2024.[155] Based on the survey results, this augmentation seems to originate less and less from “innovation” departments (one-time pools of funding) and increasingly from recurring and more permanent software budget lines, such as those of business units, R&D and product development, IT departments, etc.

[153] See for example details of a survey of US Fortune 500 enterprise leaders conducted by Andreessen and Horowitz in early 2024, elaborated in ‘16 Changes to the Way Enterprises Are Building and Buying Generative AI’, by Sarah Wang and Shangda Xu (21/03/2024), available at [147]. See also McKinsey (2024), ‘The State of AI’ and KPMG (2023), 2023 KPMG Generative AI survey.

[154] [This analysis is based on the results of a survey by PWC:](https://www.strategyand.pwc.com/de/en/industries/financial-services/embracing-the-genai-opportunity.html)
<https://www.strategyand.pwc.com/de/en/industries/financial-services/embracing-the-genai-opportunity.html>

[155] Andreessen and Horowitz (2024), *ibid.*

GenAI models: According to the same survey, enterprises seem to avoid using only one GenAI model (more than 90% of respondents indicated they are currently using at least 3 models), thereby aiming to prevent lock-in, prioritise flexibility with respect to use cases and promote readiness for new technological advancements. In addition, expectations hint at a 50%-50% future split between closed and open-source models (from 80% closed / 20% open in 2023), with over half of respondents indicating they prefer the latter. Interestingly, rather than cost-related concerns, the main reasons expressed for such an interest in open-source models are control (due to sensitive use cases and enterprise data security concerns) and the ability to customise open source models for specific use cases (using retrieval-augmented generation (RAG), fine-tuning or other methods). Last but not least, enterprises seem to prioritise internal benchmarks over external ones, irrespective of whether the fine-tuned model is closed or open source. By decreasing the importance of the advantage that closed source models typically have on publicly available benchmarks, these further levels the playing field between the closed- and open-source model competition.

In model choice and usage, cloud is still (and most probably will remain) highly influential. Self-hosting accounts for some 28% of cases. The remainder (access via API) includes a majority of access via cloud (52%) with some 38% from model providers directly. Such stickiness to cloud providers is particularly relevant in the European scenario, where the lack of large commercial sovereign Cloud Service Providers (CSPs) capable of providing their customers with access to European models highlights the risk of a long-lasting dependence on the American CSPs.

GenAI use cases and products: Most enterprises surveyed are building their own AI applications in-house, rather than buying enterprise-focused AI applications from external vendors, eased in this task by the availability of APIs and tooling for foundation models. They do this mostly for addressing familiar use cases, such as chatbots, customer support, document summarisation, software development, etc. The reasons for building applications in-house for internal use cases are easy to understand: e.g., fear of wrong or misleading responses and concerns in deploying GenAI in sensitive contexts such as health, finance, etc.[156]

[156] Noticeably, such a combination of internal development of internal use cases is bad news for companies (in particular, startups) aiming to provide corporates with similar GenAI-based apps: the already limited differentiation space gets further restricted when their potential customers prefer internal development.

Estimated GenAI adoption rates: While figures tend to vary across the literature, a 2024 McKinsey global survey reveals that 65% of those companies surveyed regularly use GenAI in at least one business unit or function, most often in marketing and sales and in product and service development, as well as in IT.[157] 67% of those surveyed expect their organisations to invest more in AI over the next three years. However, these figures look a bit different in Europe - according to the EU Digital Decade Report 2024, against a targeted AI adoption rate of 75% in enterprises by 2030, the actual figure at the end of 2023 remains quite low at only 11%, and it is unlikely the target will be achieved at this pace.[158]

GenAI impact and use in specific industrial sectors: In a survey exploring the expectations of US business executives concerning GenAI for 2024, 77% of respondents anticipated the largest impact from GenAI ahead of other emerging technologies, viewing it as a chance to gain competitive edge (64%) and boost workforce productivity (73%).[159] 71% professed plans to implement GenAI within two years, with IT/tech (56%) and operations (56%) described as priority areas, followed by marketing and sales (42%), and customer service (40%).

One might expect that GenAI would impact differently across industrial sectors, and that this would also reflect at the level of individual countries, depending on their specific industrial profiles. Focusing on companies' internal use cases, PwC classifies industrial sectors according to the impact that the adoption of GenAI is expected to have on their productivity potential.[160] This approach permits identification of the GenAI 'High Impact Accelerators' - a pool of industrial sectors, including those high in technology (e.g., pharma and bio-sciences, IT, financial services, etc.), which are poised to benefit the most from GenAI adoption. These are followed by GenAI 'Aspirants' - industries that, for a variety of reasons including a strong exposure to end users (e.g., consumer goods, retail, transport and health), while still capable of reaping important benefits, do so to a lesser extent. Finally, they refer to GenAI 'Low Potentials', comprising industries where human labour and/or technology legacy is higher, and which are expected to gain the lowest (although still significant) benefits.

[157] McKinsey (2024), *ibid*.

[158] European Commission (2024), *State of the Digital Decade*, <https://digital-strategy.ec.europa.eu/en/factpages/state-digital-decade-2024-report>

[159] KPMG (2023), *2023 KPMG Generative AI survey*, available at <https://kpmg.com/us/en/articles/2023/generative-artificial-intelligence-2023.html>

[160] PwC (2024), *Embracing the GenAI opportunity: How Europe can seize the vast potential of Generative Artificial Intelligence*, available at <https://www.strategyand.pwc.com/de/en/industries/financial-services/embracing-the-genai-opportunity.html>

GenAI high impact accelerators

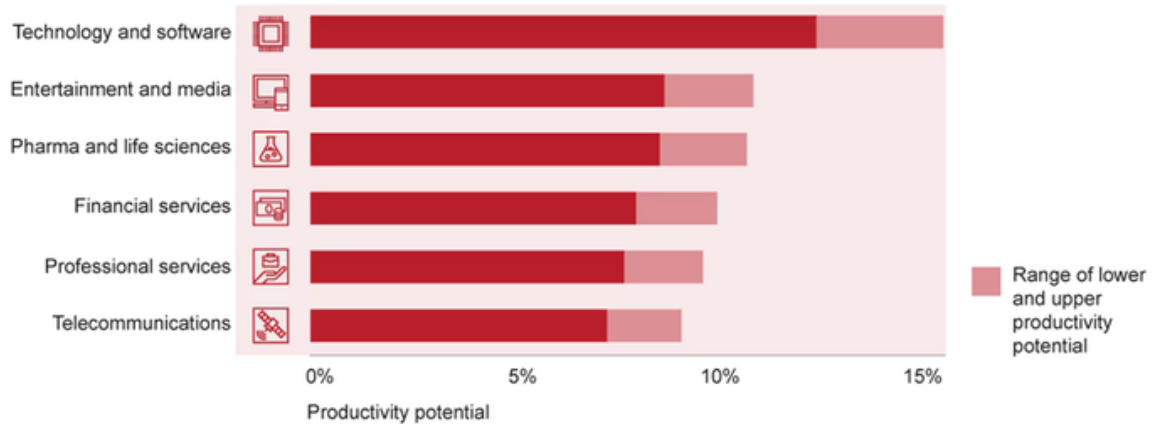


Figure 4 GenAI ‘High Impact Accelerators’ including industrial sectors which are poised to benefit the most from GenAI adoption

GenAI aspirants

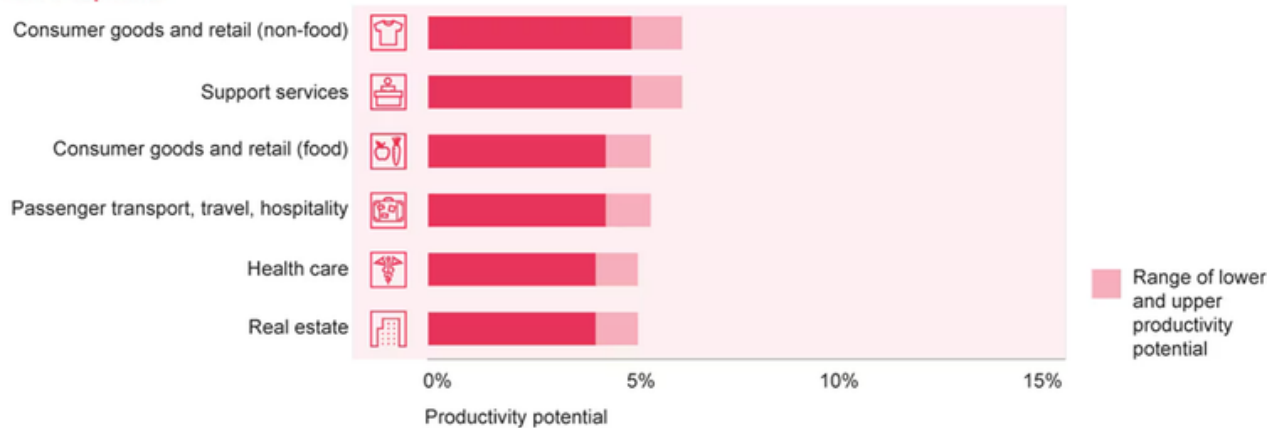


Figure 5 GenAI ‘Aspirants’ - industries that while still capable of benefiting from GenAI adoption, they do so to a lesser extent

GenAI low potentials

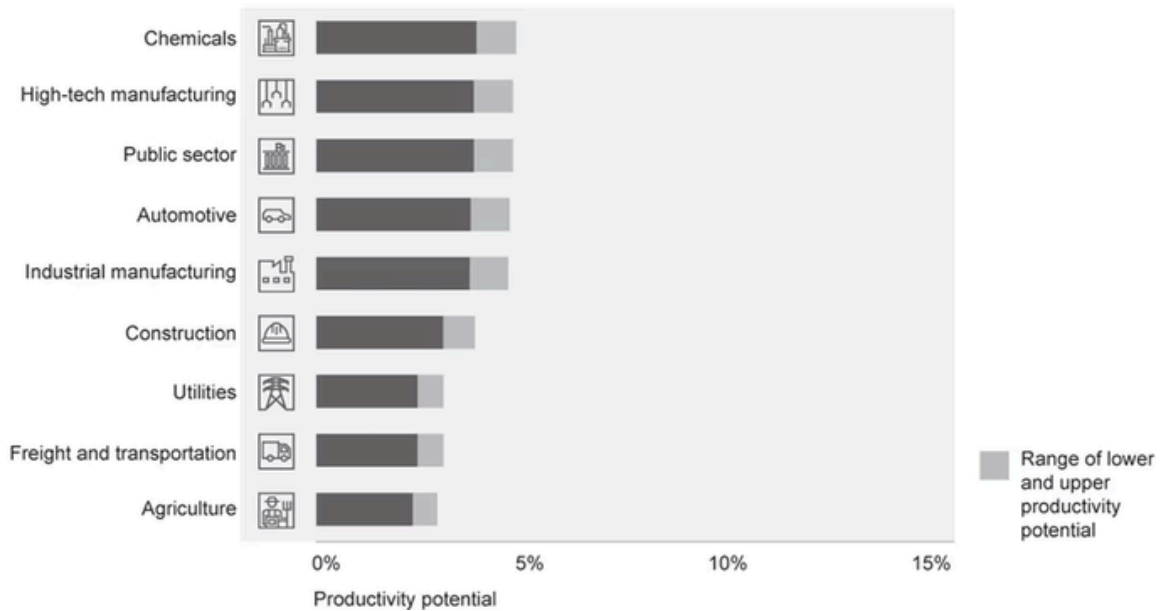


Figure 6 GenAI ‘Low Potentials’ – industries which that are expected to gain the lowest benefits due high human labour and/or technology legacy

This approach can also be extended to the country level by considering their specific industry mix, revealing that those economies with a higher presence of entertainment, life sciences, ICT, financial and professional services sectors (hence, higher in GenAI High Impact Accelerators) are poised to take stronger advantage of GenAI adoption. These are followed next by countries where the industrial mix tips the scale more towards GenAI Aspirants and finally by those with a prevalence of GenAI Low Potentials. In total, PWC estimates a 0.4% to 0.7% boost to European GDP by 2030, as a result of the adoption of GenAI.

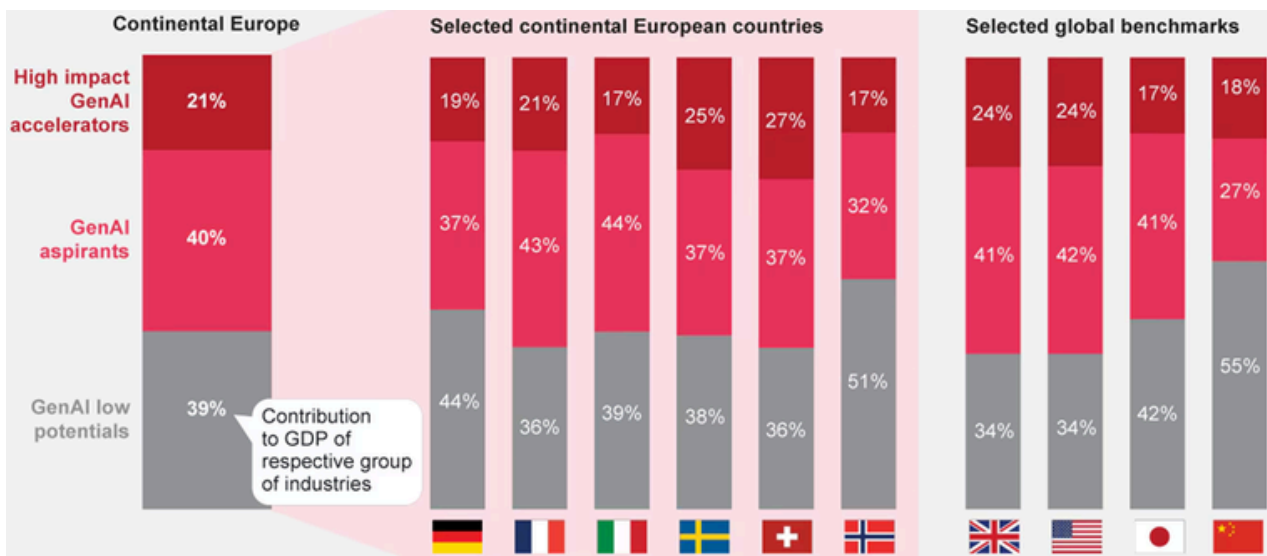


Figure 7 GenAI adoption at country level

An abstract graphic consisting of numerous thin, white, curved lines that sweep across the frame from the top left towards the bottom right. The lines are closely spaced, creating a sense of motion and depth. The background is a solid, vibrant teal color.

Generative AI and data spaces change log

Generative AI and data spaces paper change log (from v1.0 of October 2024)

Section/page	Description of change	Reason
List of contributors	Correction in the name and affiliation of one of the contributors	Name and affiliation of a contributor was incorrect in v1.0
1.2 Generative AI: an opportunity for Europe	Division of bullet points in 2 lists (1st list is driven by market analysis and figures, while 2nd list of bullet points offers further community perspectives and predictions)	To improve clarity and distinguish between market-driven and community-driven insights
2 A multidimensional perspective of the challenges of Generative AI	Adding a reference on explainability of GenAI	To provide academic support for the discussion on explainability
Pg24 "Model selection or creation"	The following sentence has been added together with a relevant reference "This is an important implementation choice and part of an ongoing discussion over the merits of the different approaches"	To reinforce the message the model selection or model creation is a fundamental choice in GenAI application development
Pg25 "Adaptation"	Some techniques for adaptation and fine-tuning have been moved to footnotes	Keep the text consistent in style and depth with the other sections
Pg28 "adaptation"	A reference to a paper recently produced by BDVA has been added	Offering information on data-quality dimensions
Pg57 "Promoting transparency and traceability"	Intro of this bullet point has been elaborated with more information considering AI Act	To strengthen the policy relevance
Pg58 2nd bullet point	Additions of more information on data governance connected to the providers of high-risk and general-purpose AI models	Enhance contextual depth and provision of practical uptake information
Pg59 3rd bullet point	Linking traceability and transparency with oversight and accountability aspects.	To connect theoretical discussion with practical implementation
Pg63 Point 1	Addition of information on which stages of maturity data spaces should be to better cope with GenAI growing needs (operational or scaling stage)	Offer more contextual depth
Pg64 Point 2	Replacing "Infrastructures developments.." with "Digital infrastructures for data sharing..."	Offer more contextual depth
Entire document	Editing / proofreading (typos, improved wording and phrasing, etc.)	Improve overall readability and clarity.

Publisher

Data Spaces Support Centre (DSSC)
c/o Fraunhofer-Gesellschaft zur Förderung
der angewandten Forschung e. V.
Hansastr. 27c
80686 Munich
Germany

Copyright

Data Spaces Support Centre, September 2024

Contact

www.dssc.eu
contact@dssc.eu

Consortium

Big Data Value Association
CapGemini Invent
FIWARE Foundation
Fraunhofer-Gesellschaft
Gaia-X
International Data Spaces Association
KU Leuven
MyData Global
SITRA
TNO
University of Galway
VTT



**DATA SPACES
SUPPORT CENTRE**



**Funded by
the European Union**

The Data Spaces Support Center receives funding from the European Union Digital Europe Programme under grant agreement n° 101083412.