DataPACT: Compliance by Design of Data/AI Operations and Pipelines*

Dumitru Roman[1], George Konstantinidis[2], Matteo Palmonari[3], Marta Musidlowska[4] and Radu Prodan[5]

[1] SINTEF AS, Forskningsveien 1, 0373 Oslo, Norway
[2] University of Southampton. Highfield, SO17 1BJ, Southampton, United Kingdom
[3] University of Milano-Bicocca, Viale Sarca 336, 20125 Milano, Italy
[4] KU Leuven Centre for IT & IP Law, Sint-Michielsstraat 6 box 3443, 3000 Leuven, Belgium
[3] University of Innsbruck, Technikerstraße 21a, 9020 Innsbruck, Austria

Abstract

DataPACT is a key initiative that develops novel tools and methodologies for efficient, compliant, ethical, and sustainable data/AI operations and pipelines. DataPACT contributes to their design, implementation, and management by embedding compliance, privacy, and environmental sustainability at their core design. It delivers compliance-by-design for data/AI operations and pipelines by developing innovative technical tools (Compliance Toolbox) and supportive methodologies (Compliance Framework) for compliance assessment and realization of data/AI pipelines designed, deployed, and executed through a set of management tools and techniques (Compliance-aware Data/AI Pipeline Toolbox). This paper presents an overview of DataPACT, focusing on motivation, methodology, and use cases.

Keywords

Data/AI operations, pipelines, compliance.

1. Introduction

In an era when ubiquitously generated and extensively utilized data drives decision-making and innovation, ensuring the compliance, fairness, and environmental sustainability of data/AI operations has become paramount. The disruptive emergence of new AI models, the increasing volume of data, the complexity and computational needs of AI systems, the interaction of different and often competing actors, and the multitude of emerging legislations pose significant compliance challenges with regulations such as the General Data Protection Regulation [1], the Data Governance Act [2], the Data Act [3], and the Artificial Intelligence Act [4]. Moreover, there is an increasing societal and business demand for ethical and transparent AI and an urgent need to mitigate the environmental impact of data and AI operations aligning with the European Green Deal's objectives [5].

Traditionally, data/AI pipeline developments neglect inherent mechanisms to ensure compliance with various legislations and ethical guidelines and commonly treat compliance as an afterthought. This oversight presents a significant challenge as the demand for compliance in data operations grows alongside the increasing complexity and scale of data and AI systems. Integrating compliance, fairness, and environmental sustainability into data/AI operations is more critical than ever, especially considering the European Green Deal's ambitious goals. Addressing the gap between existing data/AI pipeline architectures and the requirements for legal, ethical, and

* *In: Janis Grabis, Yves Wautelet, Emanuele Laurenzi, Hans-Friedrich Witschel, Peter Haase, Marco Montali, Cristina Cabanillas, Andrea Marrella, Manuel Resinas, Karolin Winter. Joint Proceedings of HybridAIMS and CAI Workshops. Co-located with CaiSE 2025.*
[1*] Corresponding author.
dumitru.roman@sintef.no (D. Roman)
0000-0001-6397-3705 (D. Roman)

environmental compliance remains an unsolved problem, underscoring the urgency and relevance of initiatives such as DataPACT. Recent developments highlight the complexity of retrofitting compliance into data/AI pipelines and the necessity for new, inherently compliant frameworks. The increased complexity associated with including AI operations significantly affects the vast numbers of small to large organizations willing to exploit AI's potential but may be overwhelmed by compliance regulations.

DataPACT is a new initiative funded through the Horizon Europe program that aims to address these challenges by developing and piloting novel tools and methodologies that enable efficient, compliant, ethical, and environmentally friendly data/AI operations and pipelines. In this context: (1) Data/AI operations involve comprehensive processes and methodologies to discover, collect, integrate, process, store, manage, and analyze data for deriving valuable insights and powering artificial intelligence (AI) systems (such operations encompass a wide range of activities, including data discovery, cleaning, linking, model training, and deployment of AI models); (2) Data/AI pipelines refer to the automated sequences of data processing and AI model training steps designed to efficiently manage data flow from its raw form to a usable state for analytics, machine learning, and AI applications (pipelines ensure scalable, repeatable, and efficient data processing, analysis, and utilization, crucial for AI systems' systematic handling and functioning).

DataPACT envisions an approach where compliance, ethics, and environmental sustainability are not afterthoughts but foundational elements of data/AI operations and pipelines. This transformation ensures a future where companies and public sector organizations can effortlessly ensure compliance with legal and ethical standards and unlock the full potential of their data assets, fostering an environment of trust and transparency for citizens. DataPACT is a holistic approach to addressing the multifaceted challenges of compliance, trust, and environmental sustainability in the data-driven world, paving the way for a more ethical, transparent, and environmentally conscious digital future. DataPACT develops innovative technical tools and tool-supported methodologies for compliance assessment and realization of data/AI pipelines designed, deployed, and executed through a set of pipeline management tools and techniques (Compliance-aware Data/AI Pipeline Toolbox).

This paper discusses provides an easy-to-understand motivating example (Section 2), outlines the DataPACT approach (Section 3), discusses relevant use cases (Section 4), and provides a summary and outlook (Section 5).

2. DataPACT motivating example

Figure 1 illustrates a typical scenario with compliance requirements addressed by DataPACT. This figure depicts a remote patient monitoring scenario (initially introduced in [6]) of an elderly patient using smart medical devices installed at home. These devices collect health-related measurements through an IoT gateway that sends the data to a cloud-based system for automated analysis. The processed data assessed against the patient's care plan sends notification alerts to the healthcare provider for further action, if necessary, and ensures timely medical intervention while enabling continuous monitoring outside traditional healthcare facilities.

A typical implementation of such a system (lower part of the figure) follows a structured data pipeline with three key steps. First, data generation, preprocessing, and routing occur at the edge (patient's home). Devices collect sensor data, format it, and push it into an MQTT broker using C programming. Second, the data storage and analytics component, implemented in JavaScript in this case, receives and processes the data, checking it against patient care plans and storing relevant records in a database. Third, the end-user application logic, also written in JavaScript and deployed in the cloud, triggers notifications to healthcare personnel when detecting anomalies or urgent health issues.

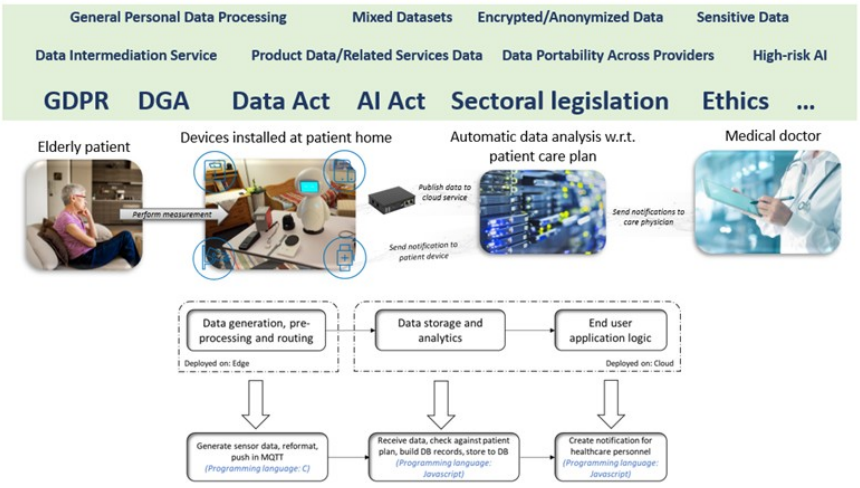

Figure 1: DataPACT remote patient monitoring example.

A challenge in deploying such a remote patient monitoring system is ensuring regulatory compliance with multiple legal frameworks and ethical standards (upper part of the figure). The scenario highlights compliance requirements such as GDPR, the AI Act, the Data Act, and sectoral legislation. These frameworks govern patient data collection, processing, storage, sharing, and security by particularly involving sensitive health information and potentially high-risk AI applications. Since this system involves automated decision-making and patient data processing, strict compliance with these regulations is essential. The GDPR plays a fundamental role in lawfully processing patient data using appropriate consent mechanisms and data protection measures. The Data Act ensures data portability and interoperability, critical for integrating patient data across healthcare providers and service platforms while preventing data monopolization. It is particularly relevant in health data-sharing scenarios, where patients can securely control and transfer their medical data. Additionally, the Data Governance Act (DGA) regulates how personal and non-personal data is shared, particularly in cases where intermediary services are involved. The AI Act is crucial in this scenario because it regulates high-risk AI applications in healthcare. If the automated data analysis system influences patient care decisions, it may fall under high-risk AI, requiring strict transparency, explainability, and human oversight. The system must ensure bias-free AI models, robust validation procedures, and clear risk assessments compliant with the AI Act. Additionally, ethical considerations must ensure that automated health alerts do not replace human medical judgment but assist healthcare providers in decision-making. Furthermore, sectoral legislation specific to healthcare may impose additional compliance requirements that prohibit patient discrimination or data misuse by insurers or employers. Encryption and anonymization of stored and transmitted data are necessary to comply with data security and privacy laws. The remote patient monitoring solution can be legally compliant, secure, and beneficial to patients and healthcare providers by designing the system with privacy-by-design principles, regulatory alignment, and ethical safeguards.

3. DataPACT methodology

DataPACT develops methods to support the compliance needs of data and AI operations and pipelines. It relies on primitive, compliance-aware data and AI operations and an infrastructure for combining them in executable analysis data/AI pipelines that achieve business goals and create value for organizations. Primitive operations and pipelines interact by design with socio-technical tools for compliance assessment and realization. Operation APIs introduced as wrappers on top of existing operations support the execution and transition to compliance-ensuring data/AI pipelines. DataPACT thus offers capabilities for the design, management, execution, and monitoring of data/AI pipelines that consider relevant compliance aspects. The DataPACT pipeline lifecycle management toolbox consists of conceptual design and runtime phases where algorithms or methods can either assess or enforce compliance aspects of data and AI operations or entire

pipelines, e.g., privacy-preservation, explainability, consent management, trustworthiness legality, fairness, transparency, energy-efficiency, adherence to (smart) contracts, and others. Assessment tools are an integral part of the methodology to support the interpretation of the pipeline and its result or certification of compliance by reusing and creating a machine-processable vocabulary for different legislations and guidelines.

DataPACT core elements. DataPACT operates with three elements: data, data/AI operations, and data/AI pipelines combined to realize business goals. DataPACT offers tools and techniques for data, operations, and pipelines to enable their management and operations compliant with GDPR, AI Act, Data Act, and Data Governance Act. Big Data in DataPACT (primarily volume, variety, and velocity) can be highly unstructured or structured. Examples of data operations include data discovery, profiling, access, cleaning, linking, validation, and structuring. Examples of AI operations include algorithm selection, feature engineering, hyperparameter tuning, and model training. Data/AI pipelines are combinations of such operations applied to data.
CEURART uses the Libertinus fonts. You may have to install these fonts on your computer. The text below shows how to locally install them.

3.1. Conceptual architecture

Typical data/AI pipelines managed by DataPACT involve a feedback sequence of steps separated by design (Identify, define, assess, realize, deploy) and runtime (execute, monitor, reassess, realize) involving a set of relevant stakeholders, depicted in Figure 2.

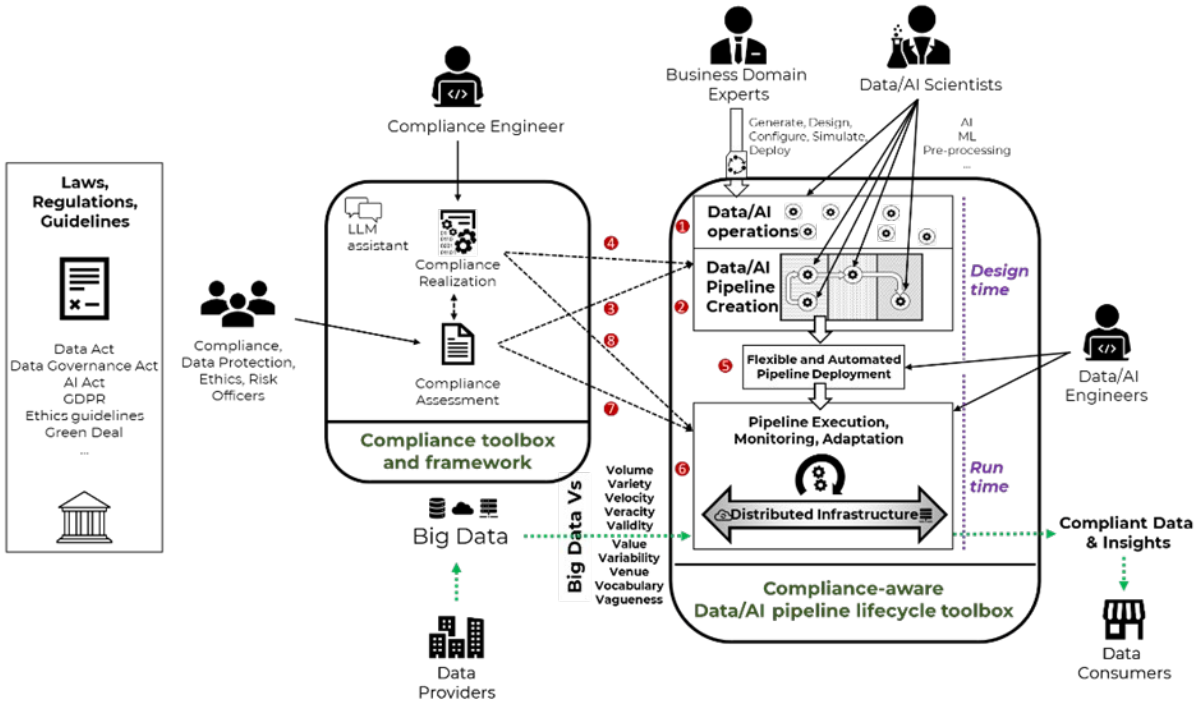

Figure 2: DataPACT methodology for compliant data/AI pipelines.

1. *Identify data/AI operations:* To perform business operations, business domain experts use domain-specific knowledge to identify relevant data and conduct operations. Data/AI scientists with AI and ML expertise specify implementation details of the operations, such as analytical models and operation-specific code. Special-purpose visual tools can add compliance-related annotations to the datasets and operations.
2. *Define data/AI pipelines:* During pipeline definition, business domain experts and data/AI scientists use the data processing requirements to structure, configure, design, and simulate pipelines, assisted by AI tools, like LLM-based generation from high-level textual specifications.

3. *Assess compliance of data/AI operations/pipelines:* After designing the pipelines, compliance, data protection, ethics, and risk officers analyze the pipelines in the light of the triggered regulations and respective guidelines (GDPR, AI Act, Data Act) through the legal and ethical assessment framework. LLM-assisted tools extract relevant information from the target regulations and guidelines, recommending actions to make the pipeline compliant.
4. *Realize data/AI operations/pipeline compliance:* Compliance engineers get recommendations for the assessment process from the previous step and implement compliance recommendations in pipeline design, possibly refining or changing them through the DataPACT tools.
5. *Deploy data/AI pipelines:* After ensuring pipeline compliance in the previous step, the data/AI engineers will provide the hardware/software infrastructure for automatically deploying and executing the pipeline using the DataPACT tools.
6. *Execute data/AI operations:* After deploying the pipeline, its execution starts with the input data from the selected data providers, depicted with the dashed arrows at the bottom of the figure. DataPACT provides the necessary tools for ensuring scalable and secure execution of the pipeline.
7. *Monitor data/AI pipelines:* Monitoring the pipeline during the execution triggers compliance reassessment if the execution results in data or steps are uncompliant.
8. *Reassess compliance of data/AI operations/pipelines:* In case of an uncompliant pipeline at run time, the Compliance, Data Protection, Ethics, and Risk Officers initiate a compliance assessment process similar to step (3) analyzing the execution of the pipeline against the target regulations, guides and issuing guidelines for ensuring compliance at runtime.
9. *Realize compliance of data/AI operations/pipelines:* Similar to step 4, compliance engineers implement the step 8 recommendations by changing or adapting pipeline execution. This step ensures that execution results (data and insights) comply with target regulations and guidelines in data consumers' interest.

3.2. DataPACT toolbox, framework and sandbox

DataPACT supports these through technical tools and methodologies for compliance assessment and realization of data/AI pipelines, designed, deployed, and executed and through a set of pipeline management tools and techniques:

- *Compliance toolbox and framework* cover several technical tools relevant to compliance (e.g., privacy policies, consent management, negotiation, bias) and methodologies supporting legislation, ethics, and social impact.
- *Compliance-aware data/AI pipeline toolbox* supports several tools to manage the lifecycle of data/AI pipelines (e.g., identification of relevant operations, design, deployment). Stakeholders use such tools to build generic pipelines for flagship applications, such as using and fine-tuning foundation models or sharing data in data spaces.
- *Regulatory sandbox environments* test the envisioned toolboxes and framework under regulatory supervision and ensure appropriate safeguards before the use case deployment (e.g., healthcare, customer support). DataPACT validates the technology-supporting compliance of data/AI pipelines for eight use cases from various domains, detailed in the next section.

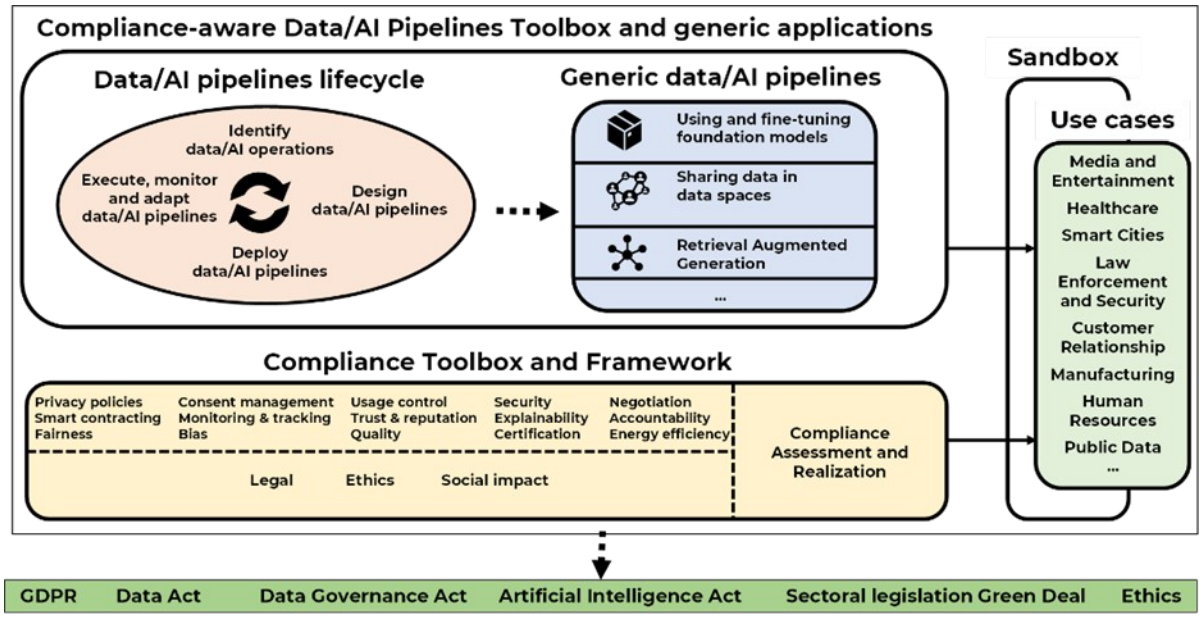

Figure 3: DataPACT toolbox, framework, sandbox, and use cases.

3.3. Artifacts and tools

Development of compliance-aware data/AI pipelines envisions:

1. Relevant, tested, and documented data/AI operations, embedding compliance-related metadata into data/AI operations and pipelines;
2. Tool for compliance-aware data/AI pipeline design with visual specification interfaces and novel LLM-assisted generation from high-level textual descriptions and requirements;
3. Tools for compliance-aware simulation, deployment, and secure execution of data/AI operations/pipelines (based on [7,8]);
4. Tool for monitoring and traceability of data, operations, and pipelines.

Privacy, consent, and access policies for data/AI pipelines envisions:

1. Fine-grained rule-based language and policy enforcement tool for specifying and enforcing access and use policies (e.g., privacy, legislation), recognizing violations in data/AI operations/pipelines, and recommending policy-compliant amendments;
2. Consent management tool for collecting, storing, and managing (e.g., updating, withdrawing) consent regarding personally identifiable data that dynamically adapts the consent-compliant execution of data/AI operations/pipelines;
3. Machine-processable contract language and tool for managing, automatically negotiating and enforcing algorithmic contracts/agreements containing detailed properties of data/AI operations and pipelines (e.g., fine-grained statements, price, energy consumption).

Trust, fairness, robustness, and explainability for compliance in data/AI pipelines envision:

1. Tool for managing trust and reputation scores of stakeholders involved in data/AI pipelines for data sharing/processing contracts and agreements;
2. Tool for the interactive inspection of biases in the data based on statistical analysis
3. Tool for the declarative specification of fairness constraints for traditional ML approaches based on a neuro-symbolic approach (based on [9]);
4. Tool for simplified assessment of fairness, robustness, and quality of LLMs, guarding against misbehaviors of generative LLMs;
5. Tool for explainability of data/AI operations and pipelines powered by LLMs.

Compliance-aware environmental sustainability of data/AI operations/pipelines envisions:

1. Tool for reporting energy consumption at different granularities of data/AI operations and pipelines and recommendation engine for suggesting energy improvements;
2. Framework for trade-off analysis between performance and energy efficiency in AI training and inference pipelines;
3. Framework for trade-off analysis between energy efficiency and privacy/fairness/robustness of data/AI pipelines.

Legal, ethical, and social impact assessment envisions:

1. Framework for legal assessment of data, operations, and pipelines based on mapping and analysis of legal provisions that trigger specific compliance requirements;
2. Framework for ethical and responsibility assessment of data, operations, and pipelines, based on assessment methodologies for human agency, diversity, societal well-being, and respect for individual rights, emphasizing non-discrimination and proper redress;
3. Sector- and use case-based methodology for social impact assessment of data/AI pipelines, engaging diverse and inclusive perspectives, involving different stakeholders, and utilizing interdisciplinary approaches to comprehend the specific impact of data/AI pipelines;
4. LLM-assisted tool for understanding compliance of data, operations, and pipelines based on natural language descriptions of legislations, regulations, ethical guidelines, and social impact;
5. Tool for legal, ethical, and environmental compliance certification of data, operations, pipelines, and associated stakeholders.

Demonstrate the usefulness in flagship complaint applications of data/AI pipelines envisions:

1. LLMOps pipeline tool for fine-tuning foundation models with built-in compliance use;
2. RAG-LLM pipeline for retrieval augmented generation improving the accuracy of contextualized foundation models (avoiding hallucinations);
3. Pipeline for supporting data-sharing workflows in data spaces.

4. DataPACT use cases

DataPACT initially selected eight use cases (UCs) to develop and validate the proposed approach, toolboxes, and framework, whereby it targeted a balanced mix of commercial (UC1, UC5, UC6, UC7) and public sector (UC2, UC3, UC4, UC8) organizations and data, and a high degree of coverage of regulations and their triggers, with high coverage of at least two principal regulations for each UC. DataPACT has selected UCs, which, due to their complexity in terms of the variety of technologies used and their innovative nature, will serve as illustrative examples to analyze the application of the various provisions of the new regulatory framework in practice and to test tools to facilitate compliance with these requirements.

Table 1 summarizes a preliminary analysis of the relevance of the selected UCs, focusing on the regulations that set the legal framework for the data economy and development of data/AI pipelines. In that context, “triggers” are contextual elements that lead to applying the legal framework when present or performing a specific function, as relevant in the context of DataPACT. Triggers indicated in the table impact the overall development and operation of the respective UCs or specific aspects of AI/data pipelines within the UCs, which require meeting requirements set by the regulation.

Table 1: DataPACT use cases.

Regulations and essential triggers		*UC1*	*UC2*	*UC3*	*UC4*	*UC5*	*UC6*	*UC7*	*UC8*
GDPR	General personal data processing	✓	✓	✓	✓	✓	✓	✓	✓
	Mixed datasets	✓	✓	✓	✓	✓	✓	✓	✓
	Encrypted anonymized data		✓		✓	✓			✓
	Sensitive data	✓	✓		✓	✓	✓	✓	
	Synthetic data	✓			✓		✓		
DGA	Public sector-protected data		✓	✓					✓
	Data intermediation			✓					
Data Act	Public sector-protected data	✓	✓	✓			✓		✓
	Smart contracts					✓			
AI Act	High-risk AI	✓	✓		✓			✓	✓
	AI with transparency obligations			✓			✓		
	General-purpose AI	✓				✓			
Social legislation		✓	✓	✓	✓	✓	✓	✓	✓
Other relevant legislation and ethics guidelines		✓	✓	✓	✓	✓	✓	✓	

- *Media and entertainment* estimate the expected impact (number of viewers and visualizations) of new media content (movies, series, advertising) using brain-computer interface technologies optimized in the cloud-edge continuum for different use scenarios.
- *Healthcare* develops a new AI decision support system to predict high-risk adverse health outcomes of patients after hospitalization and the optimal day of discharge, providing healthcare providers essential information on patients' functional status when transitioning to home care.
- *Smart city* develops a new service for compliance assessment of data pipelines, delivering data in the Urban Data Space and AI pipelines consuming data from the data space and contributing to the compliance assessment of connectors as the key component of the data space.
- *Law enforcement and security* develop a new AI solution to reduce the time and effort to analyze datasets with personal data, sensitive information, and specific watermarks for classified data and automatically determine the security classification level based on internal/national policies.
- *Customer relationship* develops an AI-based call center for customer relationship management, integrating voice and voice-to-text with customer data, marketing data, and financial metrics, performing call NLP and sentiment analysis to enhance customer support and business strategies.
- *Manufacturing*, develops a new generative AI service that improves service delivery productivity for medical imaging system devices, fined-tuned with specific knowledge and employed in a trustworthy, transparent, cost-efficient manner.
- *Human resources* develop a chatbot for employees and HR managers enquiring about personal and operative employee information, including an AI system for decision-making processes, such as promotions, monitoring and performance evaluation, and work-related relationships.
- *Public data* designs processes and pipelines for making currently restricted datasets owned and managed by municipalities available to support academic and industrial research and share data with public organizations and the private sector.

5. Summary and outlook

This paper argued the need for novel tools and methodologies that enable efficient, compliant, ethical, and sustainable data/AI operations and pipelines. We introduced DataPACT as a key initiative in this context and provided relevant challenges, a motivating example, the proposed approach, and example use cases. With a broad target group including Data/AI industries, Data/AI scientists, business experts, MLOps/DataOps, policymakers, law practitioners, compliance and data

protection officers, DataPACT addresses the lack of dedicated software tools and methodologies for supporting complaint data/AI operations and pipelines required to build compliant data-driven AI-enabled applications. DataPACT aims to deliver robust and practical software toolboxes and frameworks to support complaint data/AI pipelines, including a compliance-aware data pipelines toolbox for making data/AI pipelines compliance-aware, a compliance toolbox for auto-mating compliance-related tasks, and a compliance framework for le-gal/ethical/social assessment, validated in proper use cases products and services: media, movie making, healthcare, AI-assisted customer service, data spaces, banking, medical device services, law enforcement and security, public administration.

Acknowledgements

This work was partly funded through the Horizon Europe projects DataPACT (101189771), enRichMyData (101070284), UPCAST (101093216), and Austrian Research Promotion Agency “AIM AT Stiftungsprofessur für Edge AI” (909989). The authors acknowledge the contributions of all DataPACT project partners in developing the project.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT in order to partly: Grammar and spelling check, Paraphrase and reword. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication’s content.

References

[1] General Data Protection Regulation, Regulation - 2016/679 - EN - GDPR - EUR-Lex, https://eur-lex.europa.eu/eli/reg/2016/679/oj (2016).

[2] Data Governance Act, Regulation - 2022/868 - EN - EUR-Lex, https://eur-lex.europa.eu/eli/reg/2022/868/oj/eng (2022).

[3] Data Act, Regulation - EU - 2023/2854 - EN - EUR-Lex, https://eur-lex.europa.eu/eli/reg/2023/2854/oj/eng (2023).

[4] Artificial Intelligence Act, Regulation - EU - 2024/1689 - EN - EUR-Lex, https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng (2024).

[5] The European Green Deal, EUR-Lex - 52019DC0640 - EN - EUR-Lex, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52019DC0640 (2019).

[6] N. Nikolov, A. Solberg, R. Prodan, A. Soylu, M. Matskin, D. Roman, Container-based data pipelines on the computing continuum for remote patient monitoring. in Computer, vol. 56, no. 10, pp. 40-48, Oct. 2023, doi: 10.1109/MC.2023.3285414.

[7] D. Roman, R. Prodan, N. Nikolov, A. Soylu, M. Mataskin, A. Marella, Big Data Pipelines on the Computing Continuum: Tapping the Dark Data," in Computer, vol. 55, no. 11, pp. 74-84, Nov. 2022, doi: 10.1109/MC.2022.3154148.

[8] A. Thomas, N. Nikolov, A. Pultier, D. Roman, B. Elvesaeter, A. Soylu, SIM-PIPE DryRunner: An approach for testing container-based big data pipelines and generating simulation data". 2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC), Los Alamitos, CA, USA, 2022, pp. 1159-1164, doi: 10.1109/COMPSAC54236.2022.00182.

[9] G. Greco, M. Palmonari, Declarative Encoding of Fairness in Logic Tensor Networks. 26th European Conference on Artificial Intelligence, ECAI 2023. IOS Press, 2023, pp. 908-915. doi:10.3233/FAIA230360.