Automated Compliance for Data and AI Pipelines: the DataPACT Project

Christopher Maidens[1], Paolo Pareti[1], Dumitru Roman[2] and George Konstantinidis[1]

[1]University of Southampton, Southampton, United Kingdom
[2]SINTEF AS, Oslo, Norway

Abstract

This paper provides an overview of the EU-funded DataPACT project. This project develops a novel architecture and suite of tools to embed compliance, ethics, privacy, and environmental sustainability directly into the design and operation of data and AI pipelines. By shifting from a reactive to a "compliance by design" approach, DataPACT enables organisations to meet regulatory requirements such as GDPR and the AI Act more efficiently, while also addressing fairness, transparency, and environmental impact. Its architecture is validated through seven diverse industrial and public sector use cases, providing a foundation for trusted, sustainable, and citizen-centric European data spaces.

Keywords

Data marketplace, Negotiation, Privacy and usage control, Sustainable computing, AI legislation, AI pipelines

1. Introduction

DataPACT [1] is an EU funded project that brings together 18 partners from 16 countries, uniting academia, industry, and public organizations with the goal of ensuring trustworthy, compliant, and sustainable AI operations. This is achieved by developing tools and frameworks that embed compliance, ethics, privacy, and environmental sustainability into data and AI pipelines from the ground up. It validates its solutions across seven real-world use cases spanning industries such as healthcare, smart cities, media, and law enforcement.

The project's stated high-level objectives [2] aim to achieve key outcomes over the period of 2025 to 2027. First, it seeks to enable both **companies** and the **public sector** to **easily comply** with current and upcoming regulations, such as GDPR and the Artificial Intelligence Act, while simultaneously helping them **unlock the value of their data assets**. A second goal is to **boost citizens' confidence** that their personal data is being used in a **fair, unbiased, and compliant** manner, with their privacy and other rights respected, especially as more of our lives move online. The project also focuses on **defining, quantifying, and measuring bias in data sets**, particularly those used for AI development. Ultimately, these efforts are intended to **reduce time-to-market and development costs** for data solutions, contribute to the creation of **open, trusted, and federated Common European Data Spaces**, and even **reduce the environmental footprint** of data operations, helping to meet the Green Deal's target of no net greenhouse gas emissions by 2050.

DataPACT envisions a future where compliance, ethics, and environmental sustainability are integral to data and AI operations. This future includes simpler integration of compliance by design principles from the outset of development [3]. By seamlessly integrating compliance into data pipelines, it aims to help organizations unlock the full potential of their data while safeguarding fundamental rights and fostering trust and transparency. Thus, supporting important principles such as data sovereignty within the developing Data Space market.

The DataPACT project aims to deliver three key resources:

- A *Compliance Toolbox* that offers innovative technical solutions for assessing and ensuring compliance with regulations, privacy, and ethical guidelines.

RuleML+RR'25: Companion Proceedings of the 9th International Joint Conference on Rules and Reasoning, September 22–24, 2025, İstanbul, Türkİye

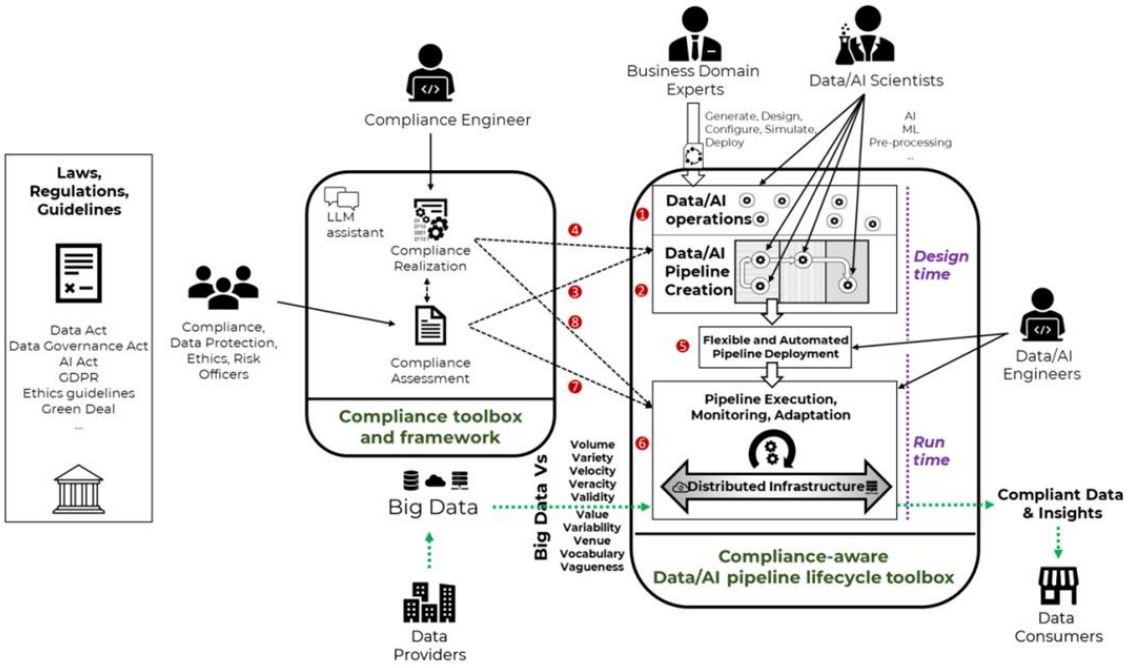

Figure 1: DataPACT Conceptual Pipeline.

- A *Compliance Framework*, which provides a methodology supported by tools to integrate compliance (and its assessment) into data and AI pipelines from the design phase.
- A *Data/AI Pipeline Toolbox*, that enables design, deployment, and execution of compliant, privacy-preserving, and sustainable data and AI pipelines.

Together, these will form a comprehensive, foundational and transformative development ecosystem, providing user-friendly tools to ensure quality and sustainability in data and AI operations in various use cases [4] [5]. Through the Compliance Toolbox and Compliance Framework, DataPACT will ensure that compliance, privacy and ethics are embedded throughout the design and execution of data pipelines. The Compliance-aware Data/AI Pipeline Toolbox will simplify pipeline management, enabling straightforward, cost-effective regulatory adherence while promoting fairness, privacy, and environmental responsibility. Supporting a focus on the creation of open, federated Common European Data Spaces, DataPACT will enhance data interoperability, traceability, and reliability across sectors, providing a foundation for trusted, data-driven solutions. [6]

2. Architectural Summary

DataPACT will contribute to both the design and operation of compliant data/AI pipelines as illustrated in Figure 1. This figure illustrates the two major phases of consideration. That is the design and implementation phase (the left hand box) and the operational phase (the right hand box). It shows how the DataPACT outcomes will enhance and support these phases. It also highlights how the operational phase may need to revisit the design and implementation phase and refine the pipeline if non compliance is observed during operation. This following text provides a little more detail on these concepts.

The **design and implementation of data and AI pipelines** involves a series of steps. The first phase is to **identify the specific data and AI operations** that are required for a project. During this step, DataPACT will develop specialized visual tools that can be used to add compliance-related annotations to the datasets and operations. Next, it will **define the data and AI pipelines**. This will

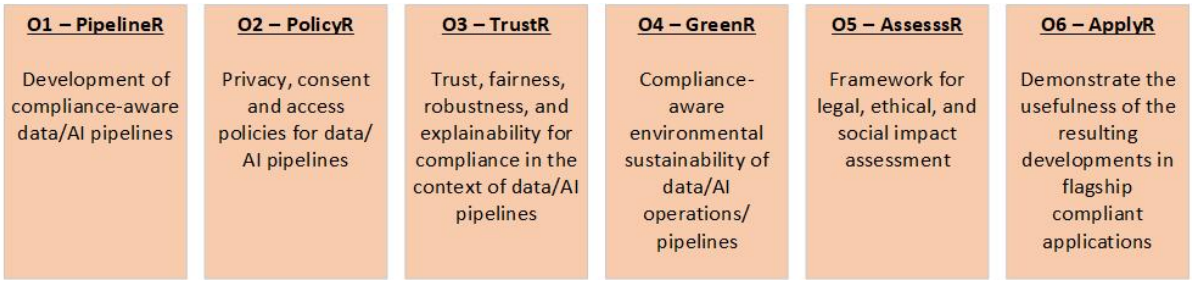

Figure 2: Main objectives of the DataPACT project.

be done using AI tools, such as large language models (LLMs), that can be leveraged to automatically generate pipelines from high-level, textual requirements.

Once the pipelines are defined, the next crucial step is to **assess their compliance**. DataPACT will develop LLM-assisted tools that will extract relevant information from target regulations and guidelines, providing a clear set of recommended actions to make the pipeline compliant. These recommendations are then passed on to a **Compliance Engineer**, who implements the changes and refines the pipeline's design using the project's tools. Finally, with compliance ensured, the data and AI engineers can **deploy the pipelines** by setting up the necessary hardware and software infrastructure for automated execution.

Once a data or AI pipeline is designed, the next phase is **operation**, which focuses on execution, monitoring, and compliance. The project provides the necessary tools to ensure the pipeline's execution is both scalable and secure. As the pipeline runs, it is continuously **monitored for compliance** and the project also provides tools to enhance automated capabilities. Should the execution produce data or steps that make it non-compliant, a **compliance reassessment is automatically triggered**. This reassessment involves **efficiently** redeploying the design tools to reset and reconfigure the pipeline, ensuring it meets all regulatory standards before continuing.

3. Objectives

To support this conceptual vision a number of major supporting objectives are being actioned. These are summarized in Figure 2.

PipelineR will develop tooling and approaches for the **embedding of compliance-related metadata** directly into data and AI operations and pipelines. It will develop a **compliance-aware pipeline design tool**, which combines visual specification interfaces with innovative LLM-assisted generation from high-level textual descriptions. It will also provide tools for **compliance-aware simulation, deployment, and secure execution** of pipelines. Finally, it includes **monitoring and traceability tools** to track data, operations, and pipelines throughout their lifecycle, ensuring continued adherence to compliance standards.

PolicyR provides a set of tools designed to manage and enforce compliance within data and AI operations. It includes a **fine-grained, rule-based language and policy enforcement tool** for specifying and enforcing access and usage policies related to privacy and legislation. This tool can also identify violations in data and AI operations and pipelines, offering recommendations for policy-compliant amendments. Additionally, PolicyR features a **consent management tool** that handles the collection, storage, and management (e.g., updating, withdrawing) of consent for personally identifiable data, dynamically adapting the execution of data and AI pipelines to remain compliant with user consent. Lastly, the platform offers a **machine-processable contract language and tool** for the management, automatic negotiation, and enforcement of algorithmic contracts and agreements. These contracts can include detailed properties of data and AI operations, such as fine-grained statements, pricing, and energy consumption.

TrustR provides a suite of tools for ensuring the integrity, fairness, and transparency of data and AI operations. It includes a tool for **managing the trust and reputation scores** of stakeholders involved in data-sharing and processing agreements. To ensure fairness, TrustR offers a tool for the **interactive inspection of biases in data**, utilizing statistical analysis. It also provides a tool for the **declarative specification of fairness constraints** for traditional machine learning models through a neuro-symbolic approach. For generative AI, there's a tool for the **simplified assessment of the fairness, robustness, and quality of LLMs**, which helps guard against misbehaviors. Finally, the platform features a tool for the **explainability of data and AI operations and pipelines**, powered by LLMs, to provide insights into how decisions are made.

GreenR focuses on the promotion of environmental sustainability in data and AI operations. It features a tool for **reporting energy consumption** at various levels of granularity within data and AI operations and pipelines, along with a recommendation engine that suggests **energy improvements**. It will also provide a framework for performing **trade-off analysis** between performance and energy efficiency in AI training and inference pipelines. Furthermore, it offers a framework to analyze the trade-off between energy efficiency and other critical factors such as **privacy, fairness, and robustness** in data and AI pipelines.

AssessR provides a comprehensive suite of tools for evaluating data, operations, and pipelines from multiple perspectives. It offers frameworks for both **legal assessment** and **ethical and responsibility assessment**. Additionally, it includes a **sector- and use case-based methodology for social impact assessment**. To simplify the process of understanding compliance, AssessR features an **LLM-assisted tool** that interprets natural language descriptions of regulations, ethical guidelines, and social impacts. This allows for a clearer understanding of compliance requirements. Finally, the platform provides a **certification tool** to formally certify the legal, ethical, and environmental compliance of data, operations, pipelines, and the stakeholders involved.

ApplyR offers a suite of tools for the development and deployment of compliant AI systems and data-sharing workflows. It includes an **LLMOps pipeline tool** specifically for fine-tuning foundation models with compliance built-in. To enhance the accuracy and reliability of contextualized foundation models, ApplyR features a **RAG-LLM pipeline for Retrieval Augmented Generation**, which helps prevent hallucinations. Additionally, the platform provides a **pipeline to support data sharing workflows** within various data spaces.

Where appropriate DataPACT will make use of state-of-the-art tooling recently developed within relevant EU programmes and projects aimed at developing Data Spaces and their capabilities across the EU. An example of this is the EU Funded project UPCAST [7], which developed several technologies relevant to the **PolicyR** objective. A set of tools (in this case developed by University of Southampton) have been developed to better support policy definition and negotiation between data providers. These include **policy editors and supporting management services** allowing fine-grained definition of data privacy policies using the machine-understandable ODRL [8] policy definition language. A **user-friendly, web-based interface** that allows non-technical users to easily create, update, store, and share ODRL policies. To enable a more precise vocabulary for policy creation, the platform supports the **integration of domain-specific ontologies**. Another relevant tool that will be reused and expanded is a **Negotiation and Contracting service** that facilitates **seamless and policy-compliant data transactions between providers and consumers** by managing a complete **IDSA compliant negotiation lifecycle** [9] from initiation to contract finalization. The system addresses inherent differences in processing intentions, usage constraints, and pricing expectations through a structured, machine-readable contract framework built on standards such as ODRL and DPV [10].

4. Validation

DataPACT validates its core results through a strong selection of representative use cases offered by SMEs, large companies, and public sector organizations in relevant areas, including media and entertainment, healthcare, smart cities, law enforcement and security, customer relationship management,

manufacturing, and public data. The seven use cases of DataPACT are the following.

UC1 (Media and Entertainment) develops a new product for estimating the expected impact (number of viewers and visualizations) of new media content (movies, series, advertising) by using brain-computer interfaces technologies optimised in the cloud-edge continuum for different use scenarios.

UC2 (Healthcare) develops a new AI decision support system to predict high-risk negative health outcomes of patients after hospitalization and the optimal day of discharge, providing healthcare providers essential information on patients' functional status when transitioning from hospital to home care.

UC3 (Smart Cities) 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.

UC4 (Law Enforcement and Security) develops a new AI solution to reduce the time and effort needed to analyse documents/datasets with personal data, sensitive information, and specific watermarks for classified data and automatically determine the security classification level based on internal/national policies.

UC5 (Customer Relationship) develops an AI-based Call Center for Customer Relationship Management, integrating voice, voice-to-text with customer data, marketing data and financial metrics, performing call NLP and sentiment analysis to enhance customer support and business strategies.

UC6 (Manufacturing) develops a GenAI solution that improves service delivery productivity for medical imaging system devices, fine-tuned with specific knowledge and employed in a trustworthy, transparent, cost-efficient manner.

UC7 (Public Data) develops a new service for designing 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 sector organisations and the private sector.

5. Conclusions

The EU DataPACT project aims to develop frameworks and tools that will improve how data and AI systems are built by promoting a *compliance by design* approach. This shifts the process from a traditional, reactive model to one where compliance is a foundational element of data and AI operations. The project offers **holistic and proactive compliance**, providing a framework and toolbox that simplifies a process often seen as an expensive afterthought, which is particularly crucial given the growing complexity of regulations like GDPR and the AI Act. This scope is **broad**, extending beyond just legal compliance to include ethical, environmental, and trust-related aspects. For example, tools like GreenR and TrustR are designed to measure environmental impact and assess bias. By creating tools that ensure fairness, transparency, and respect for privacy, the project also aims to **empower citizens and foster trust** in data-driven systems, a vital, non-economic value for the long-term health of the European data economy. Furthermore, DataPACT directly contributes to the EU's digital strategy by **enabling the European data economy** through the creation of open, trusted, and federated data spaces, which will lower barriers to data sharing and collaboration. Additionally, the project's delivery of **practical and actionable tools**—such as PipelineR and PolicyR—moves beyond theoretical frameworks, offering concrete solutions for legal assessments, automated contract negotiation, and workflow streamlining, ultimately helping organizations reduce costs. The comprehensive range of Use Cases and wide stakeholder engagement ensures a thorough validation and dynamic communication of best practice and practical implementations during the lifetime of this development.

Acknowledgments

This work was funded by the UKRI Horizon Europe guarantee funding scheme for the Horizon Europe project DataPACT (10.3030/101189771).

Declaration on Generative AI

During the preparation of this work, the author(s) used Google Gemini in order to: 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] European Union, Ai-driven data operations and compliance technologies (ai, data and robotics partnership) (ia), 2023. URL: https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/topic-details/horizon-cl4-2024-data-01-01, dataPACT Funding Call.

[2] DataPACT, Datapact propsoal, 2024. URL: https://sintef.sharepoint.com, dataPACT Proposal.

[3] D. Roman, Compliance for data/ai pipelines, 2025. URL: https://datapact.se-sto-1.linodeobjects.com/Data_PACT_Data_Week2025_Dumitru_Roman_c3bda85bbb.pdf, compliance for Data/AI Pipelines - Presentation.

[4] SINTEF, Datapact compliance by design of data/ai operations and pipelines, 2025. URL: https://www.sintef.no/en/projects/2025/datapact/, sINTEF DataPACT Project Website.

[5] O. Galanets, Elevit symposium athens digital health week 2025 - the role of ai and multivariate data in enhancing transitional care european health data space: Legislative framework, 2025. URL: https://datapact.se-sto-1.linodeobjects.com/ELEVIT_Symposium_Athens_Digital_Health_Week2025_Olga_Galanets_IDSA_2535586ddc.pdf, iDSA Presentation.

[6] DataPACT Project Team, Datapact project website, 2025. URL: https://datapact.eu/, dataPACT Project Website.

[7] UPCAST, Upcast project home, 2023. URL: https://www.upcast-project.eu/, uPCAST Project Home.

[8] W3C, Odrl information model 2.2 w3c recommendation 15 february 2018, 2018. URL: https://www.w3.org/TR/odrl-model/, oDRL Information Model 2.2.

[9] IDSA, Layers of the reference architecture model 3.4.3 contract negotiation, 2025. URL: https://docs.internationaldataspaces.org/ids-knowledgebase/ids-ram-4/layers-of-the-reference-architecture-model/3-layers-of-the-reference-architecture-model/3_4_process_layer/3_4_3_contract_negotiation, iDSA Contract Negotiation Flow.

[10] W3C, Data privacy vocabulary (dpv) version 2.1, 2025. URL: https://w3c.github.io/dpv/2.1/dpv/, data Privacy Vocabulary (DPV) version 2.1.