

A Testing and Experimentation Facility Architecture for Trustworthy AI Systems in the Energy Sector

Evangelos Karakolis¹, Spiros Mouzakitis¹, Alexandros Tzortzis¹, Panagiota Rempi¹,
Stavroula Chalvantzi¹, Theodosios Pountridis¹, Sotiris Pelekis¹, Dimitris Askounis¹

¹School of Electrical and Computer Engineering,
National Technical University of Athens,
Athens, Greece

{vkarakolis, smouzakitis, atzortzis, prempi, schalvantzi, tpountridis, spelekis, askous}@epu.ntua.gr

Abstract—The rapid adoption of Artificial Intelligence (AI) in the energy sector raises significant challenges related to trustworthiness, regulatory compliance, scalability, and interoperability. While AI-based energy solutions are increasingly deployed across power grids, buildings, and distributed energy systems, there is a lack of unified infrastructures that enable systematic testing, validation, and experimentation under realistic conditions. This paper presents the EnergyGuard architecture, a European AI Testing and Experimentation Facility (TEF) designed to support the development, validation, and acceptance of AI-driven energy solutions. EnergyGuard provides an integrated platform combining secure data management, AI development and experimentation environments, high-performance computing resources, and a dedicated Trustworthy AI acceptance layer. The architecture supports the full AI lifecycle from initial experimentation and trustworthiness assessment to model deployment and continuous monitoring, enabling AI systems to adapt over time in response to evolving operational conditions, data drift and emerging compliance requirements. The architecture is aligned with major European digital initiatives, including data spaces, AI-on-Demand, and emerging regulatory frameworks such as the EU AI Act. Rather than focusing on algorithmic performance, this paper contributes a reference architecture that operationalizes trustworthy-by-design principles and enables scalable, policy-aligned AI experimentation for the energy domain. Specifically, the paper investigates how data management, experimentation, computation, and trustworthiness assessment can be provided by a single platform, so as to operationalize trustworthy-by-design AI development for the energy sector.

Index Terms—AI Testing and Experimentation Facility, Trustworthy AI, Energy Systems, EU AI Act, Digital Twins, High-Performance Computing, Data Spaces, MLOps

I. INTRODUCTION

A. Context, Motivation and Related Work

Artificial Intelligence (AI) is increasingly used across the energy sector to support forecasting, optimization, anomaly detection, and decision support in power grids, buildings, microgrids, and emerging energy communities [1]. Recent advances extend beyond predictive analytics to include Digital Twins (DTs) for simulation and monitoring of energy systems [2], as well as Large Language Models (LLMs) for decision support in complex energy-related settings [3]. These tech-

nologies are central to managing the growing complexity and digitalization of modern energy infrastructures.

Despite promising results, the deployment of AI systems in operational energy environments remains challenging. Energy infrastructures are critical assets with strict requirements in reliability, safety, and regulatory compliance, due to their significant financial, environmental, and societal implications [4]. Within the European regulatory landscape, frameworks such as the AI Act [5] and the Assessment List for Trustworthy AI (ALTAI) [6] further reinforce these obligations by requiring trustworthiness through robustness, transparency, accountability, and human oversight. In practice, however, current AI development in the energy domain often emphasizes model performance under controlled experimental conditions, with limited support for large-scale validation, stress testing under realistic scenarios, and structured trustworthiness assessments prior to deployment. As a result, a gap persists between research prototypes and market-ready AI solutions.

Several European initiatives have addressed individual aspects of the technical, trustworthy and governance challenges associated with deploying AI in the energy sector. The Horizon 2020 project I-ENERGY [7] promoted AI-based energy services through a federated data-driven innovation ecosystem, delivering modular, secure analytics tools [8] for applications such as forecasting [9]–[11] and demand response, and building energy retrofits [12], [13]. The project also emphasized FAIR-compliant data management, semantic modeling, and explainability mechanisms to improve transparency and interoperability. At the infrastructure level, GAIA-X [14] introduced a federated and trustworthy data framework, supporting secure and sovereign data sharing through standardized identity management, policy-based usage control, and compliance across distributed cloud and edge environments [15]. Similarly, the International Data Spaces Reference Architecture Model (IDS-RAM) [16] provides a structured architectural approach for secure and policy-enforced data exchange, enabling organizations to maintain control over data access and usage through certified roles and interoperable connectors. Building on these principles within the energy sector, the Common European Energy Data Space (CEEDS) [17] aims to establish an interoperable ecosystem for energy data exchange at the European level, which is compliant with EU legislation and incorporates

This work is supported by the EnergyGuard project, co-funded by the European Union under the Horizon Europe Programme.

semantic standards, trust mechanisms, and governance models.

To address the need for unified environments to support the safe, compliant, and scalable development of AI solutions, the European Union has introduced the concept of AI Testing and Experimentation Facilities (TEFs) [18]. Domain-specific TEFs enable testing, validating and scaling of trustworthy AI systems under realistic conditions, bridging experimental solutions and operational deployment.

B. Objectives and Research Question

This paper presents the EnergyGuard architecture, a TEF tailored to the energy sector. EnergyGuard is an end-to-end experimentation platform that supports AI development, digital twin-based testing, high-performance experimentation, and structured trustworthiness and compliance assessment. In doing so, it directly addresses the challenge of transforming AI prototypes into robust, scalable, and trustworthy real-world systems for a safety-critical domain. The paper is guided by the following research question: how can a unified testing and experimentation facility be architected to operationalize trustworthy-by-design AI development for the energy sector, integrating data, experimentation, computation, and compliance assessment within a single environment? In answering this, the paper makes three contributions: (1) it identifies a set of design drivers derived from SME and startup needs, regulatory expectations, and concrete operational scenarios; (2) it presents the EnergyGuard reference architecture as a concrete realisation of those drivers; and (3) it provides an architectural validation through two representative energy-sector use cases. Accordingly, the paper focuses on architectural design choices, component interactions, and alignment with the EU digital ecosystem, rather than on algorithmic performance.

II. ARCHITECTURAL CONTEXT AND DESIGN DRIVERS

The EnergyGuard architecture was developed based on a concrete set of requirements rather than from abstract principles alone. These requirements were elicited from four sources: the functional and technical specifications of the EnergyGuard project; consultations with SMEs and startups as the platform's intended users; regulatory expectations arising from the EU AI Act and the ALTAI framework; and a set of representative operational scenarios in the energy domain. From these sources we distilled five design drivers, presented below, which collectively shape the architecture. The use cases of Section V then serve as architectural validation, examining whether the resulting design supports realistic AI testing and experimentation workflows.

The EnergyGuard architecture builds upon existing European initiatives and reference architectures addressing AI experimentation, data sharing, and digital sovereignty in the energy sector. Initiatives such as I-ENERGY, GAIA-X, IDSRAM, and CEEDS have introduced important concepts such as federated data access, modular AI toolchains, and sector-specific data spaces, which form a necessary foundation for large-scale AI adoption. Despite this progress, existing efforts typically address isolated stages of the AI lifecycle and lack a

unified architectural approach that integrates experimentation, trustworthiness assessment, regulatory alignment, and scalable execution within a single facility.

The proposed consolidated architecture aims to address real needs of AI solution providers in the energy sector. In this context, we identify five operational scenarios that reflect such needs posed by SMEs and startups, seeking to bring AI-driven energy services from a prototype to market-ready version, under the European regulatory framework.

Compliance and trustworthiness evaluation. AI service providers in energy need to systematically assess whether their solutions meet trustworthiness and regulatory requirements prior to market entry. EU AI Act classifies certain energy-related AI systems as high-risk, which raises further the requirements for robustness, transparency and human oversight. EnergyGuard acceptance environment provides trustworthiness assessment workflows, AI ethics tools, and security assessment to produce verifiable compliance artifacts. This driver is realised by the Trustworthy AI Assessment Environment (Section III-C), which combines questionnaire-based AI Act and ALTAI assessments with automated robustness, fairness, and explainability benchmarks. As the architecture's core innovation, this environment is currently under development.

Validation on realistic energy data and use cases. Before deployment to production, AI solutions must be validated against representative, high-quality data, in real operational conditions. Access to such data is often prohibitive for small providers due to data sovereignty, governance constraints and data acquisition costs. EnergyGuard secure data management infrastructure provides curated, harmonized, semantically enriched energy datasets within a governed, policy-compliant environment through data space technologies. Hence, validation can be performed on both realistic and legally accessible data. This driver maps onto the Data Management component (Section III-A) — comprising the secure data lake, the ingestion and harmonisation layer, and the data sharing layer — which is the most mature part of the platform and is already operational.

Access to cloud-based testbeds and digital twins. Realistic validation of AI solutions beyond static datasets requires dynamic interactive environments that emulate operational behavior. DTs of energy assets enable AI solutions to be tested against simulated but operationally aligned conditions, without risk to production infrastructure. To this end, EnergyGuard's AI Development and Experimentation Sandbox, exposes a variety of Energy DTs through standardised interfaces, enabling AI experiments to interact with simulated operational environments in a reproducible and controlled manner. This driver is served by the Energy DTs and standardised interfaces of the AI Development and Experimentation Sandbox (Section III-B); the supporting DT portfolio is at mixed maturity, with some twins operational and others still under development.

Scalable experimentation via high-performance computing. Training and stress-testing AI models at scale demands computational resources usually unreachable to small companies. EnergyGuard integration with MeLuxina HPC, offers GPU-accelerated execution of AI experiments and training,

abstracted behind user-friendly interfaces. This enables service providers scale experiments and validation workflows without managing HPC environments directly. This driver is addressed by the HPC Infrastructure component (Section III-D), which integrates MeLuxina through JupyterHub and Slurm-managed GPU nodes; this integration is available as beta version.

Continuous monitoring of AI behaviour over time. Trustworthy AI is not a one-time process but an ongoing commitment, since AI solutions may experience data drift, model degradation, or emergent failures under evolving operational conditions. EnergyGuard’s Acceptance environment in collaboration with the AI development sandbox, provides experiment tracking, model versioning and continuous assessment capabilities, ensuring timely identification of performance, trustworthiness and security issues that might arise over time. This driver is realised jointly by the Acceptance Environment and the Sandbox’s experiment-tracking and model-versioning capabilities; while several tracking functionalities are available, drift detection over time is not yet supported.

III. ENERGYGUARD REFERENCE ARCHITECTURE

EnergyGuard architecture (Figure 1) is designed as a modular, service-oriented architecture supporting the full AI experimentation lifecycle. The platform integrates four core backend components: data management, an AI development and experimentation sandbox, high-performance computing infrastructure, and a Trustworthy AI acceptance environment. All components are accessed through a unified user interface and are protected by centralized identity and access management. The architecture emphasizes interoperability, scalability, and compliance-by-design, enabling multiple stakeholders to experiment with AI solutions under realistic and controlled conditions. Each of these components is presented in detail in the following subsections.

A. Data Management and Interoperable Data Sharing

The Data Management component enables secure, interoperable, and policy-compliant handling of heterogeneous energy data across the entire experimentation lifecycle. Rather than acting as a simple storage subsystem, it provides a structured data backbone supporting harmonisation, semantic enrichment, analytics, and controlled data exchange among multiple stakeholders.

Energy data originate from a **data sources layer** (Energy Data Sources, Fig. 1), representing external systems, including TEFs or real-world environments, that produce raw energy data with heterogeneous formats, protocols, and sampling frequencies.

To handle heterogeneity, EnergyGuard introduces a **data ingestion and harmonisation layer**, responsible for collecting raw data streams and transforming them into a unified internal representation. This process follows source-aware extract–transform–load (ETL) pipelines aligned with a common data model. Transformations are implemented through source-specific ETL modules that are periodically triggered by a scheduler and consolidated through a data bus before storage.

Persisted data are managed within the **secure data lake layer** which separates time-series measurements from contextual and static information while maintaining synchronization and logical consistency between them. Time-series data and static or contextual data are managed using storage technologies appropriate to their access patterns and structure, including relational and graph-based systems. Data quality is an important concern and therefore, through validation and correction services, it is ensured that datasets used in digital twins and AI workflows satisfy minimum consistency requirements. In addition, a metadata catalogue captures descriptive information about available datasets to support discovery and traceability across the platform.

Efficient access to the curated data is enabled using an **intelligent querying layer**, which abstracts the underlying storage complexity from end users and applications. A unified SQL query engine supports structured access across time-series and static data enabling business intelligence and pre-analytics. This layer also includes a data enricher component, that enhances datasets with semantic and contextual information, improving their usability in AI experimentation scenarios.

Interoperability, beyond the platform’s boundaries, is addressed through the **data sharing layer**, which supports decentralised and federated exchange. The curated data are exposed in linked-data formats using JSON-LD representations and accessed through standard APIs. This approach allows external systems to consume EnergyGuard data in a machine-readable and semantically aligned manner. The architecture is designed to interoperate with data space technologies allowing controlled dataset exposure, policy-defined usage constraints, and traceability across organisational boundaries. Specifically, data providers can retain sovereignty over shared assets, including control over access conditions and permitted usage, while secure and auditable exchanges support compliance, trust, and accountability in cross-platform data sharing scenarios.

Given the sensitivity of energy-sector data, **privacy, security, and governance mechanisms** are treated as critical concerns and enforced across all layers, including access control and authorisation, encryption of data at rest and in transit, and auditing for traceability and regulatory compliance.

Finally, an **ontology and semantics layer** provides a shared conceptual model of entities and relationships within the energy domain. This layer supports semantic interoperability, contextualisation of datasets, and advanced discovery capabilities, and provides a common reference point for both AI developers and platform services.

B. AI Development and Experimentation Sandbox

The AI Development and Experimentation Sandbox is the core experimentation environment of the EnergyGuard platform. It serves as the primary workspace where users can design, execute, and evaluate AI workflows. It provides controlled access to Digital Twins (DTs), datasets and energy related applications (e.g. DeepTSF [19]), thereby enabling reproducible experimentation under realistic operational conditions.

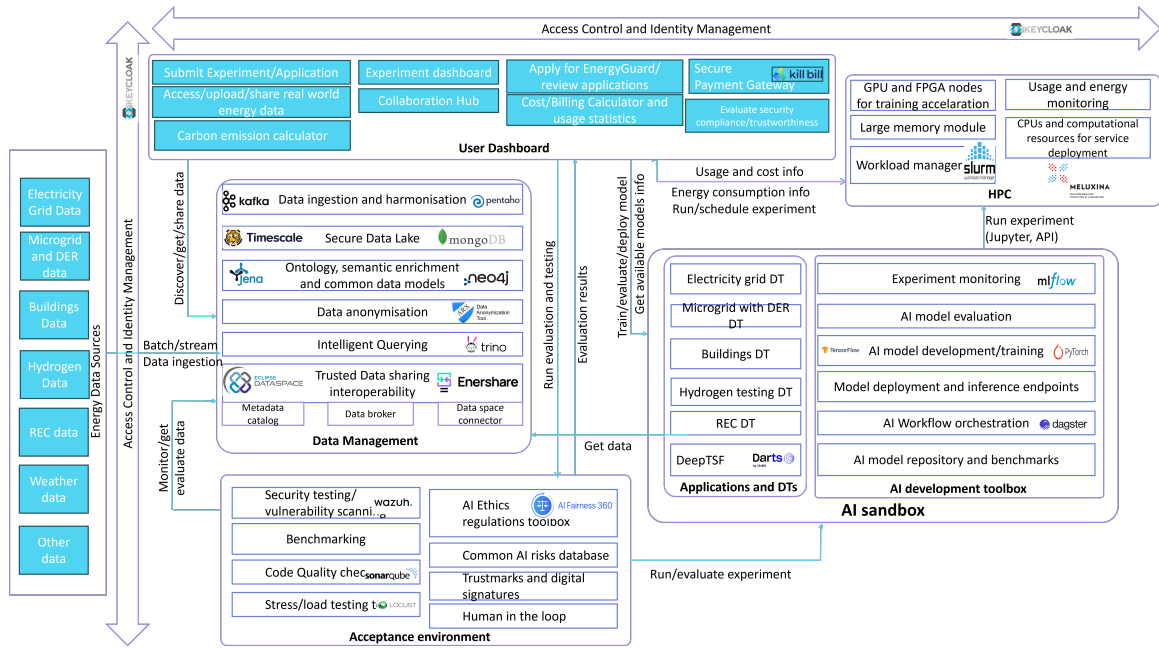


Fig. 1. EnergyGuard Reference architecture

The sandbox is organized into two groups of subcomponents: (i) **Energy Analytics Applications and DTs**, and (ii) the **AI Development Toolbox**.

1) *Energy Analytics applications and DTs*: Energy DTs in power and energy systems are commonly described as comprising a physical entity (or observational data), bidirectional communication/updates, a virtual entity, data management, and provisioned services (e.g., simulation, optimization, fault diagnosis) [20]. Recent surveys on power systems DTs emphasize data federation, interoperability challenges, and the breadth of operational use cases (e.g., grid monitoring, planning, resilience, and anomaly/fault-related analytics) [21]. EnergyGuard consists of various energy related DTs that offer a varied portfolio spanning smart grids, microgrids, hydrogen/hybrid systems, buildings, and energy communities. These DTs are exposed through RESTful interfaces so that ML experiments can call simulations and queries as external services rather than embedding these simulations in each experiment. Specifically, EnergyGuard offers the following DTs:

- **Electricity grid DT**: The Portuguese Transmission System DT uses Transmission System Operator (TSO) historical load, topology, and fault data stored in the EnergyGuard data lake to emulate grid operation in real time and enable safe AI/ML testing (e.g., forecasting, fault detection) without affecting live operations. This follows transmission DT practice, where topology/state models are combined with historical operational data for safe forecasting and contingency/fault analysis [21].
- **Microgrid with DER DT**: The distributed energy resources (DER) DT models a bounded microgrid of interconnected loads and DERs operating as a single controllable entity in grid-connected or islanded mode.

It includes local small-scale generation and storage, including export-capable storage.

- **Buildings DT**: The Riga residential DT integrates GEORIGA 3D [22], cadastral/city-infrastructure data, and real-time solar/weather feeds in a common data lake, combining heating telemetry from 5,000 buildings with audit records for high-resolution historical and live inputs.
- **Hydrogen testing DT**: Integrates data from various PEM systems, monitoring the key operational variables that govern efficiency, safety, and degradation in PEM/SOEC (Proton Exchange Membrane / Solid Oxide Electrolysis Cell) electrolyzers [23].
- **Renewable Energy Community DT**: This community scale low-voltage DT mirrors the local low voltage network of Antrodoco, Italy in near real time by combining generation, load, storage, and weather streams and exposing live/historical APIs.

EnergyGuard also offers some standalone energy related applications. DeepTSF [19] supports the development and deployment of time-series forecasts using state-of-the-art AI models: from an uploaded time series and a few user-selected options such as forecast horizon and model type, it runs a complete preprocessing, training, and evaluation pipeline automatically, building on the Darts [24] library. It is integrated with MLflow [25] for experiment tracking and model deployment and with Dagster [26] for workflow orchestration, and additionally supports hyper-parameter tuning and SHAP-based [27] explainability. While applicable to forecasting tasks generally, it is particularly suited to energy-related use cases.

2) *AI development toolbox*:

- **AI model development**: Lets users implement a repro-

ducible MLOps pipeline spanning data loading, preprocessing, training, and evaluation within isolated container environments (JupyterHub [28] single-user servers). Data is retrieved from the platform’s data lake in heterogeneous formats, together with ontology level metadata when available, and may be preprocessed using user defined transformations (imputation, missing-value handling, and anomaly detection) prior to training. Model training can run on the HPC environment, with progress exposed through APIs for long running jobs; various ML frameworks are supported. Models are evaluated using task-appropriate metrics and plots logged to MLflow [25] consistent with established MLOps practice [29].

- **Model deployment and inference endpoints:** The final model can be served in a containerised environment for inference exposed as an endpoint usable by its creator or, if made public, by other users and experiments.
- **Experiment monitoring:** All experiments conducted within the platform are tracked through MLflow [25], which provides experiment and pipeline-stage separation, a model registry, and persistent storage of metrics, plots, and metadata. Experiment metadata is held in a relational database and large artefacts such as models and preprocessed datasets in a dedicated object store.
- **AI Workflow orchestration:** EnergyGuard uses Dagster [26] to orchestrate multi-stage AI pipelines, providing ordered stage execution, live log streaming, per-stage timing for benchmarking, and scheduled runs.
- **AI model repository and benchmarks:** Models produced by users are saved to a versioned repository, alongside their evaluation results.

C. Trustworthy AI Assessment Environment

The Trustworthy AI Assessment Environment is the core innovation of EnergyGuard architecture. It provides structured evaluation of AI solutions with respect to regulatory and ethical alignment, and technical trustworthiness. Positioned after experimentation and prior to deployment, it acts as the final validation checkpoint in the EnergyGuard pipeline.

The environment follows a hybrid architecture. Questionnaire-based assessments are served directly through the platform dashboard, while automated benchmarks execute as containerized MLflow runs within the AI Sandbox, reusing the same orchestration and tracking infrastructure described in Section III-B. This design avoids duplicating compute and storage resources while maintaining logical separation between experimentation and acceptance evaluation.

1) *Questionnaire-Based Assessments:* Two structured questionnaires support qualitative compliance review.

The *AI Act Compliance Assistant* guides users through the obligations imposed on high-risk AI systems as defined by the EU AI Act, covering risk management, data governance, technical documentation, transparency, human oversight, and accuracy requirements. Given that most energy-sector AI applications are expected to fall under the high-risk category, this tool produces a gap analysis report identifying unmet require-

ments. The *ALTAI Self-Assessment* implements the Assessment List for Trustworthy AI, enabling energy sector users to evaluate their solutions against the seven trustworthiness pillars defined by the HLEG on AI. Note here that the supporting AI risk database ensures an approach that is representative of the energy sector [4]. Both questionnaires are project-scoped: results are linked to specific user projects and MLflow experiments and persisted alongside model metadata, enabling longitudinal tracking as solutions evolve.

2) *Automated Trustworthiness Benchmarks:* Three automated benchmarks complement the qualitative assessments by executing quantitative evaluations against registered MLflow models. Each benchmark is packaged as a containerised pipeline triggered from the dashboard and executed within the Sandbox environment. It should be noted that these benchmarks build on established, community-maintained tools, the Adversarial Robustness Toolbox, AI Fairness 360, and SHAP, rather than introducing new algorithms; the contribution of EnergyGuard lies in their integration into a unified acceptance workflow and their configuration for the regression and forecasting tasks characteristic of the energy sector.

Adversarial Robustness. These benchmarks evaluate model resilience to input perturbations and data poisoning strategies, implemented through frameworks such as ART [30] and adapted for regression and forecasting tasks that are common in the energy sector. Prediction degradation under adversarial conditions is measured and evaluated against configurable tolerance thresholds, generating a robustness score.

Fairness and Bias. Fairness in energy AI differs from its more familiar framing in domains such as lending or hiring, and warrants explicit definition. In the EnergyGuard context, fairness concerns whether an AI system delivers equitable accuracy across consumer segments and building types, provides fair access to services such as demand-response programmes and retrofit investment recommendations, and avoids systematically disadvantaging particular regions, network areas, or categories of energy user. A forecasting model that is accurate in aggregate may still perform markedly worse for specific consumer groups or low-voltage areas; the fairness benchmarks are intended to surface precisely such disparities. Fairness benchmarks assess whether model predictions exhibit statistically significant disparities across sensitive attributes, using tools like AI Fairness 360 [31]. Group fairness metrics like disparate impact, equalised odds, and demographic parity are computed and evaluated against configurable thresholds. Particular attention is paid to energy-specific fairness concerns like equitable access to demand response programmes and retrofit investment recommendations.

Explainability. These benchmarks audit the quality and reliability of model explanations using tools such as SHAP. Beyond generating feature attribution reports, the benchmarks assess metrics such as explanation faithfulness (whether masking high-attribution features degrades predictions proportionally), stability (consistency of explanations under small input perturbations), and coverage (proportion of predictions with sufficiently informative explanations). For time-series models,

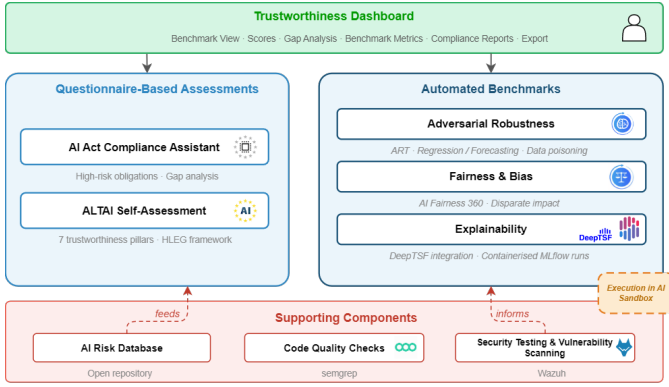


Fig. 2. Architecture of the Trustworthy AI Assessment Environment.

the benchmarks integrate with DeepTSF’s existing SHAP pipeline to ensure consistency between experiment-phase and acceptance-phase explanations.

3) *Supporting Components*: The environment is complemented by three cross-cutting components. The already mentioned *AI Risk Database*, maintained as an open repository, which catalogs known risks for the energy sector and recommended mitigations mapped to ALTAI categories and AI Act obligations, populated through pilot feedback and partner contributions. *Security testing and vulnerability scanning*, leveraging tools such as Wazuh [32], evaluates AI service endpoints against adversarial network conditions and known CVEs. *Code quality checks* via semgrep [33] assess software maturity and flag security anti-patterns prior to deployment.

All assessment results including questionnaire scores, benchmark metrics, security findings, and code quality reports are aggregated into a unified trustworthiness dashboard that supports filtering by project, benchmark type, and assessment status, and enables export of compliance reports for potential regulatory or investor use. The architecture of the Trustworthy AI Assessment Environment is illustrated in Figure 2.

D. High-Performance Computing Infrastructure

EnergyGuard integrates LuxProvide’s MeluXina HPC infrastructure [34] to support computationally intensive AI workflows, including large-scale model training, hyperparameter tuning and stress testing. The architecture follows a modular design in which persistent orchestration services (e.g., model serving, experiment tracking) are hosted on cloud-based Virtual Machines (VMs), through an OpenStack-based cloud module [35], while compute-intensive workloads are dispatched to multi-GPU nodes managed by Slurm [36] workflow scheduler. This separation enables stable platform operation alongside scalable high-performance execution.

Within this environment, EnergyGuard deploys a JupyterHub [28] instance as a unified entry point for HPC-supported AI development. JupyterHub abstracts Slurm allocation procedures for end users, providing on-demand access to GPU nodes with workload isolation across users and datasets. Software execution on GPU nodes runs through Apptainer-based containerised runtimes [37] rather than user-managed

installations, ensuring reproducibility, multi-user isolation, and compliance with infrastructure constraints.

MeluXina’s planned extension with a dedicated Kubernetes partition alongside Slurm is expected to improve support for AI pipeline orchestration and enable controlled public ingress. These changes will be abstracted through a common execution layer, allowing EnergyGuard workflows to remain stable as the underlying infrastructure evolves.

IV. ALIGNMENT WITH THE EUROPEAN DIGITAL ECOSYSTEM

EnergyGuard has been designed to interoperate with the European digital ecosystem rather than operate in isolation. This is achieved through alignment with the following initiatives.

A. AI on Demand Platform (AIoD)

At the level of AI asset access and experimentation, EnergyGuard aligns with the objectives and architectural principles of AIoD [38]. AIoD aims to provide unified access to AI assets, experimentation capabilities, and deployment mechanisms for the European research community. This alignment is reflected in EnergyGuard’s support for isolated experimentation environments, structured access to AI models and workflows, and metadata-driven discovery of AI resources. The adoption of compatible architectural patterns allows potential interoperability with AIoD services while preserving EnergyGuard’s role as a domain-specific TEF for energy applications. EnergyGuard AI assets will be published to the AIoD platform for reuse across Europe, and alignment with, and publication to the DOME Marketplace [39] will be examined.

B. European Energy Data Spaces

Beyond AI asset experimentation, trusted and sovereign data sharing is a core requirement for EnergyGuard, motivating close alignment with European data space architectures [40]. EnergyGuard adopts design principles consistent with initiatives such as GAIA-X and IDS-RAM [41], [42], emphasising decentralisation and policy-controlled access for data exchange beyond the platform. In practice, the architecture supports controlled data exposure, metadata-based discovery, federated access, and the enforcement of data usage policies across organisations, enabling participants to exchange energy-related data while retaining control over access and usage conditions. Compatibility with sector-specific initiatives such as ENERSHARE [43], further ensures alignment with emerging standards for energy data spaces.

C. Complementary European Initiatives

EnergyGuard’s design also considers alignment with European Digital Innovation Hubs (EDIHs) [44] and emerging AI Factories. EDIHs act as regional intermediaries facilitating SME access to digital technologies; EnergyGuard can extend their reach by providing remote access to AI experimentation environments, DTs, curated energy datasets, and trustworthiness assessment capabilities not available locally. EnergyGuard is also complementary to AI Factories: whereas

the latter target large-scale AI development and operational scaling, EnergyGuard emphasises controlled pre-deployment testing, validation, and trustworthiness assessment. Having been funded alongside other domain-specific TEFs, EnergyGuard further pursues collaboration and interoperability through comprehensive API catalogues to avoid fragmented solutions across the EU ecosystem.

V. USE-CASE-DRIVEN ARCHITECTURAL VALIDATION

To validate the practical relevance and completeness of the proposed architecture, we present two representative use cases. Rather than experimental evaluations, these serve as architectural instantiations demonstrating how the platform supports realistic AI testing and experimentation workflows within a TEF context, and were chosen to exercise different architectural components, including data management, AI experimentation, HPC, and trustworthiness assessment.

A. AI-driven decision support for energy efficiency investments in residential buildings

One representative use case involves the Energy Efficiency Investments Decision Support System (EFI-DSS), an AI-driven application designed to support cost-effective renovation planning for multi-apartment residential buildings in Riga, Latvia [45]. The tool generates virtual energy performance certificates and estimates the expected impact of alternative retrofit strategies based on limited user input. In this context, the EnergyGuard TEF provides a controlled experimentation environment for the systematic testing, validation, and enhancement of the application, in alignment with emerging regulatory requirements.

The use case exercises all core architectural components. Enriched building and energy-performance datasets in the Data Lake are accessed through the Data Management component to retrain EFI-DSS's predictive models on a broader, more representative data foundation, while Riga's Buildings DT supplies continuously updated audit and consumption data for validation under realistic conditions. HPC resources enable scalable training and computationally intensive extensions, including LLM integration for natural-language interaction and automated generation of tailored retrofit packages. Functional and model enhancements (e.g., coverage of additional building types and new impact indicators like emission reductions) are developed within the AI Development and Experimentation Sandbox, and explainability, robustness, and regulatory-compliance criteria are assessed within the Trustworthy AI Assessment Environment. The application is thus validated not only technically but also from a trustworthiness perspective, demonstrating end-to-end experimentation support for an existing AI-based decision support system.

B. AI-driven forecasting of the Portuguese Transmission System's energy consumption and generation for grid analysis and market outcome validation

This use case aims to improve operational planning for the Portuguese transmission system by forecasting electricity

consumption and renewable generation under realistic conditions and assessing their impact through a Digital Twin (DT) of the grid. Recent operational time-series data is used to train forecasting models producing week-ahead load and generation predictions for all grid nodes; these forecasts are then fed into the DT to run grid analysis under the forecasted operating conditions. Combined with market data, this analysis validates whether day-ahead market results remain feasible from a grid perspective or risk congestion and other constraint violations. Both deep learning models (e.g. NBEATS [46], RNNs [47]) and machine learning models (e.g. LightGBM [48]) can support such forecasting, while foundation models such as Chronos-2 [49] offer a promising alternative with potential to be fine-tuned on the DT's historical data.

The use case validates the EnergyGuard architecture by exercising its core assets in an integrated, modular workflow. JupyterHub [28] provides the HPC-backed environment for training and evaluating foundation models, while DeepTSF [19] supports the remaining forecasting tasks through a complete ML pipeline. The DT of the Portuguese Transmission Network provides the grid simulation and analysis, accessible through JupyterHub and also available as a standalone service for applications such as fault forecasting and related grid analytics.

VI. DISCUSSION

A. Answering Research Questions

The EnergyGuard architecture demonstrates how AI experimentation, trustworthiness assessment, and regulatory alignment can be integrated within a single infrastructure, but doing so involves deliberate design trade-offs. The originality of EnergyGuard does not lie in any individual component (data spaces, digital twins, AI sandboxing, HPC, MLOps tooling, and trustworthy-AI assessment instruments). Rather, the contribution is the integration of these elements into a single, coherent reference architecture for an energy-sector TEF, in which each stage of the AI lifecycle, from data access through experimentation and computation to compliance assessment and continuous monitoring, is addressed within one governed environment.

A primary challenge is balancing platform sophistication with usability. EnergyGuard provides a variety of AI services, yet must remain accessible to SMEs and startups without deep infrastructure expertise. In this context, we focus on providing HPC resources through JupyterHub, shielding users from Slurm scheduling and resource allocation complexities while preserving access to GPU-accelerated computation. Similarly, the platform unifies training, trustworthiness assessment, security evaluation and deployment monitoring within a single environment, avoiding the fragmentation that arises when each AI model lifecycle stage requires a separate toolchain.

A second trade-off concerns centralisation versus federation. EnergyGuard operates simultaneously as a centralised platform, providing a portfolio of services under unified access control, and as a federated participant in data space ecosystems that supports peer to peer data exchange. Reconciling these

two modes, particularly with respect to access policy enforcement and auditability is an ongoing architectural challenge.

A primary limitation is that the architecture is presented and validated at the design level; empirical validation through live pilot deployments remains future work. Other limitations center on data space interoperability and federated experimentation. The data space landscape remains fragmented, with multiple connector technologies that are not mutually compatible, and the overhead of deploying a compliant connector is discouraging for small organisations. A planned improvement is to abstract this complexity through technology-agnostic one-click connector provisioning. On the experimentation side, cross-team, cross-organisation workflows that require data to remain local are not yet supported. Future work will explore federated learning and differential privacy techniques to enable such scenarios.

Several lessons emerged during the architectural design process. First, data space interoperability proved more difficult than anticipated: the proliferation of mutually incompatible connector technologies meant that "alignment with data spaces" could not be treated as a solved integration step, but became a design concern in its own right. Second, shielding users from infrastructure complexity, particularly HPC scheduling, was repeatedly found to be as important to adoption by SMEs as the analytical capabilities themselves. Third, the need for unified access control across all EnergyGuard components has proved very challenging, given that commercial offerings like MeLuxina or DTs, have their own authentication mechanisms stemming from different requirements than EnergyGuard. Harmonising these into a unified authentication and authorisation mechanism required considerable design effort.

B. Match and Contribution

The paper aligns directly with the ICE 2026 focus on digital transformation and, in particular, with the special session on trustworthy, adaptive AI systems. EnergyGuard addresses the central concern of that session — transforming AI prototypes into robust, scalable, and trustworthy real-world systems — by providing the architectural infrastructure through which energy-sector AI solutions can be systematically tested, assessed for trustworthiness, and prepared for compliant deployment. The architecture's support for continuous monitoring and re-evaluation further speaks to the "adaptive" dimension of the session theme, since it allows AI systems to be reassessed as operational conditions and regulatory expectations evolve. The work also addresses the engineering- and technology-management objectives of IEEE TEMS. It concerns the management of an emerging technology (trustworthy AI) within a regulated, safety-critical industry, and translates abstract regulatory expectations such as the EU AI Act and ALTAI into a practical framework of architectural components and assessment workflows. It engages explicitly with implementation challenges, including data space fragmentation and the centralisation–federation tension, and addresses value creation

by lowering the barrier for SMEs to move energy-AI solutions from prototype to market at lower cost and risk.

VII. CONCLUSION AND FUTURE WORK

This paper presented the EnergyGuard architecture, an AI Testing and Experimentation Facility tailored to the energy sector. By combining AI development tools, digital twins, high-performance computing, and trustworthiness assessment within a unified platform, EnergyGuard enables systematic, scalable, and policy-aligned AI experimentation, supporting the full AI lifecycle. The architecture is aligned with the European digital ecosystem, including data spaces, AIoD, and the EU AI Act, and provides a reference blueprint for future AI experimentation infrastructures in safety-critical domains.

Future work will focus on three priorities: empirical validation of the architecture through live pilot deployments, on the platform's DTs and partner facilities; extension of the continuous monitoring capabilities to support automated drift detection, and triggered reevaluation workflows; and development of technology-agnostic data space connector provisioning to lower the barrier for federated data exchange for SMEs. Together, these steps will consolidate EnergyGuard's role as a reference infrastructure for scalable, policy-aligned, and adaptive AI experimentation in safety-critical domains.

ACKNOWLEDGMENT

This work has been funded by the European Union's Horizon Europe Research and Innovation programme under the EnergyGuard project, grant agreement No. 101172705. The sole responsibility for the content of this paper lies with the authors; the paper does not necessarily reflect the opinion of the European Commission.

REFERENCES

- [1] X. Wang, H. Wang, B. Bhandari, and L. Cheng, "Ai-empowered methods for smart energy consumption: A review of load forecasting, anomaly detection and demand response," *International Journal of Precision Engineering and Manufacturing-Green Technology*, vol. 11, pp. 1–31, 09 2023.
- [2] A. Fuller, Z. Fan, C. Day, and C. Barlow, "Digital twin: Enabling technologies, challenges and open research," *IEEE Access*, vol. 8, pp. 108 952–108 971, 2020.
- [3] G. Antonesi, T. Cioara, I. Anghel, V. Michalakopoulos, E. Sarvas, and L. Todorean, "A systematic review of transformers and large language models in the energy sector: towards agentic digital twins," *Applied Energy*, vol. 401, p. 126670, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S030626192501400X>
- [4] S. Pelekis, E. Karakolis, and et al., "Trustworthy artificial intelligence in the energy sector: Landscape analysis and evaluation framework," in *2024 IEEE International Conference on Engineering, Technology, and Innovation (ICE/ITMC)*, 2024, pp. 1–10.
- [5] C. Directorate-General for Communications Networks and E. C. Technology, "Ai act — regulatory framework on artificial intelligence," Web page, 2025, last update: 8 October 2025. [Online]. Available: <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>
- [6] —, "The assessment list for trustworthy artificial intelligence (ALTAI) for self-assessment," Web page, European Commission, Tech. Rep., July 2019, last update: 8 March 2021. [Online]. Available: <https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment>
- [7] E. Karakolis, S. Pelekis, and et al., "Artificial intelligence for next generation energy services across europe—the i-nergy project," in *International Conferences e-Society 2022 and Mobile Learning 2022*, 2022, pp. 61–68.

- [8] S. Mouzakitis, O. Markaki, and et al., "Enhancing decision support systems for the energy sector with sustainable artificial intelligence solutions," in *SAI Intelligent Systems Conference*, vol. 823 LNNS. Springer, Cham, 2024, pp. 61–70. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-47724-9_5
- [9] S. Pelekis, E. Karakolis, and et al., "In search of deep learning architectures for load forecasting: A comparative analysis and the impact of the covid-19 pandemic on model performance," in *2022 13th International Conference on Information, Intelligence, Systems & Applications (IISA)*. IEEE, 7 2022, pp. 1–8. [Online]. Available: <https://ieeexplore.ieee.org/document/9904363/>
- [10] S. Pelekis, I.-K. Seisopoulos, and et al., "A comparative assessment of deep learning models for day-ahead load forecasting: Investigating key accuracy drivers," *Sustainable Energy, Grids and Networks*, vol. 36, p. 101171, 12 2023. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S2352467723001790>
- [11] A. M. Tzortzis, S. Pelekis, and et al., "Transfer learning for day-ahead load forecasting: A case study on european national electricity demand time series," *Mathematics* 2024, Vol. 12, Page 19, vol. 12, p. 19, 12 2023. [Online]. Available: <https://www.mdpi.com/2227-7390/12/11/19/htm>
- [12] V. Michalakopoulos, S. Pelekis, and et al., "Efi-dss: A machine learning-based decision support system for energy efficiency investments in buildings," *2025 10th International Conference on Smart and Sustainable Technologies, SpliTech* 2025, 2025. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/11091625>
- [13] P. Rempi, S. Pelekis, and et al., "A trustworthy by design classification model for building energy retrofit decision support," *arXiv preprint arXiv:2504.06055*, 12 2025. [Online]. Available: <http://arxiv.org/abs/2504.06055>
- [14] "GAIA-X Project," accessed: 2026-02-20. [Online]. Available: <https://gaia-x.eu>
- [15] E. Karakolis, S. Pelekis, and et al., "The i-nergy reference architecture for the provision of next generation energy services through artificial intelligence," in *International Conferences e-Society 2023 and Mobile Learning 2023*, 2023, pp. 95–102.
- [16] B. Otto, S. Steinbuss, A. Teuscher, and S. Lohmann, "International data spaces association reference architecture model version 3.0," INTERNATIONAL DATA SPACES ASSOCIATION (IDSA), Tech. Rep., 2019, accessed: 2026-02-20. [Online]. Available: <https://internationaldataspaces.org/wp-content/uploads/IDS-RAM-3.0-2019.pdf>
- [17] A. Dognini, A. Monti, and et al., "Blueprint of the common european energy data space," Jul. 2024. [Online]. Available: <https://doi.org/10.5281/zenodo.12635358>
- [18] European Commission, "Sectorial AI Testing and Experimentation Facilities under the Digital Europe Programme," last update: 21 May 2025. [Online]. Available: <https://digital-strategy.ec.europa.eu/en/policies/testing-and-experimentation-facilities>
- [19] S. Pelekis, T. Pountridis, and et al., "Deeptsf: Codeless machine learning operations for time series forecasting," *SoftwareX*, vol. 27, p. 101758, 9 2024.
- [20] J. B. Heluany and V. Gkioulos, "A review on digital twins for power generation and distribution," *International Journal of Information Security*, vol. 23, pp. 1171–1195, 4 2024.
- [21] M. M. Thwe, A. Ştefanov, V. S. Rajkumar, and P. Palensky, "Digital twins for power systems: Review of current practices, requirements, enabling technologies, data federation, and challenges," *IEEE Access*, vol. 13, pp. 105 517–105 540, 2025.
- [22] "Georiga homepage," accessed: 2026-02-25. [Online]. Available: <https://georiga.eu/en/3d-riga>
- [23] C. R. Wang, J. M. Stansberry, and et al., "Proton exchange membrane (pem) water electrolysis: Cell-level considerations for gigawatt-scale deployment," *Chemical Reviews*, vol. 125, pp. 1257–1302, 2 2025.
- [24] J. Herzen, F. Lässig, and et al., "Darts: User-friendly modern machine learning for time series," *Journal of Machine Learning Research*, vol. 23, pp. 1–6, 10 2021.
- [25] "Mlflow homepage," accessed: 2026-02-25. [Online]. Available: <https://mlflow.org>
- [26] "Dagster homepage," accessed: 2026-02-25. [Online]. Available: <https://dagster.io>
- [27] S. M. Lundberg, P. G. Allen, and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017. [Online]. Available: <https://github.com/slundberg/shap>
- [28] "Jupyterhub homepage," accessed: 2026-02-25. [Online]. Available: <https://jupyter.org/hub>
- [29] D. Kreuzberger, N. Köhl, and S. Hirschl, "Machine learning operations (mlops): Overview, definition, and architecture," *IEEE Access*, vol. 11, pp. 31 866–31 879, 2023.
- [30] M.-I. Nicolae, M. Sinn, M. N. Tran, B. Buesser, A. Rawat, M. Wistuba, V. Zantedeschi, N. Baracaldo, B. Chen, H. Ludwig, I. Molloy, and B. Edwards, "Adversarial robustness toolbox v1.0.0," *arXiv preprint arXiv:1807.01069*, 2018, <https://github.com/Trusted-AI/adversarial-robustness-toolbox>.
- [31] R. K. E. Bellamy, K. Dey, and et al., "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 4:1–4:15, 2019.
- [32] "Wazuh – open source security platform," <https://wazuh.com>, 2025, accessed: 2026-02-25.
- [33] semgrep, "Lightweight static analysis for many languages." 2020. [Online]. Available: <https://github.com/semgrep/semgrep>
- [34] "Meluxina homepage," accessed: 2026-02-25. [Online]. Available: <https://www.luxprovide.lu/meluxina>
- [35] "Openstack homepage," accessed: 2026-02-25. [Online]. Available: <https://www.openstack.org>
- [36] "Slurm homepage," accessed: 2026-02-25. [Online]. Available: <https://slurm.schedmd.com/overview.html>
- [37] "Apptainer homepage," accessed: 2026-02-25. [Online]. Available: <https://apptainer.org>
- [38] A. Troumpoukis, A. Marinakis, and et al., "Deployai: Ai-on-demand platform and marketplace for industrial applications within europe," in *Artificial Intelligence Applications and Innovations. AIAI 2025 IFIP WG 12.5 International Workshops*. Cham: Springer Nature Switzerland, 2025, pp. 115–127.
- [39] K. Bialas, S. Sengupta, and et al., "Enhancing cloud marketplace operations with infrastructure as code: Dome - a case study," in *2024 IEEE Globecom Workshops (GC Wkshps)*. IEEE, Dec. 2024, p. 1–6. [Online]. Available: <http://dx.doi.org/10.1109/GCWkshp64532.2024.11100474>
- [40] M. Gourié, H. Barancourt, M. Boust, P. Calvez, M. Laskowski, A.-S. Taillandier, L. Tilman, M. Uslar, and O. Warweg, *The Energy Data Space: The Path to a European Approach for Energy*. Springer International Publishing, 2022, p. 535–575. [Online]. Available: http://dx.doi.org/10.1007/978-3-030-93975-5_33
- [41] H. Tardieu, *Role of Gaia-X in the European Data Space Ecosystem*. Springer International Publishing, 2022, p. 41–59. [Online]. Available: http://dx.doi.org/10.1007/978-3-030-93975-5_4
- [42] B. Otto, *The Evolution of Data Spaces*. Springer International Publishing, 2022, p. 3–15. [Online]. Available: http://dx.doi.org/10.1007/978-3-030-93975-5_1
- [43] S. Jimenez, "Interoperability framework in energy data spaces," 2023. [Online]. Available: <https://zenodo.org/doi/10.5281/zenodo.10117882>
- [44] E. Carpentier, D. D'Adda, D. Nepelski, and J. Stake, *European Digital Innovation Hubs Network's activities and customers*. Luxembourg (Luxembourg): Publications Office of the European Union, 2025, no. KJ-01-25-198-EN-N (online). [Online]. Available: <https://data.europa.eu/doi/10.2760/7784020>
- [45] V. Michalakopoulos, S. Pelekis, and et al., "Efi-dss: A machine learning-based decision support system for energy efficiency investments in buildings," in *2025 10th International Conference on Smart and Sustainable Technologies (SpliTech)*, 2025, pp. 1–7.
- [46] B. N. Oreshkin, G. Dudek, P. Pełka, and E. Turkina, "N-beats neural network for mid-term electricity load forecasting," *Applied Energy*, vol. 293, p. 116918, 7 2021.
- [47] L. Sehovac and K. Grolinger, "Deep learning for load forecasting: Sequence to sequence recurrent neural networks with attention," *IEEE Access*, vol. 8, pp. 36 411–36 426, 2020.
- [48] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems* 30, 2017. [Online]. Available: <https://github.com/Microsoft/LightGBM>
- [49] A. F. Ansari, O. Shchur, and et al., "Chronos-2: From univariate to universal forecasting," *arXiv preprint arXiv:2510.15821*, 10 2025. [Online]. Available: <http://arxiv.org/abs/2510.15821>