Project „Misijomis grįstų mokslo ir inovacijų programų įgyvendinimas“ (Project Nr. 02-002-P-0001) report

Project name: ***CTI-balanced***

Responsible/Implementation partner (-s): ***MRU***

Action result: ***COMPLIANCE FRAMEWORK FOR MISSION 2 AI SYSTEMS UNDER THE EU AI ACT AND HARMONISED STANDARDS***

Table of Contents

Executive Summary *4*

Introduction *5*

1. AI System Descriptions and Risk Overview *6*

1.1. AI System Description: Espionage Prevention Monitor with Detection Capabilities (EPMwDC). 9

1.2. AI System Description: HIPSTer – Hybrid, Information, Psychological, Societal Threats Handling System 11

1.3. AI System Description: AICP-FIMI – AI Driven Cloud Platform to Counter FIMI 14

1.4. AI System Description: CTI-balanced – Cyber Threat Intelligence-based Sectorial and National Collective Cybersecurity Balanced Incentives System 16

1.5. AI System Description: OSOTS – Open Source Operational Technology Sensor 18

2. The Digital Omnibus on AI – Key Pillars and Strategic Implications for MISSION 2 Projects 20

1. Pillar 1: Simplifications to AI Act Obligations and Conformity Procedures 20

2. Pillar 2: Privacy and Data Protection Enhancements 20

3. Pillar 3: Cybersecurity and System Robustness 20

4. Pillar 4: Innovation Support and Proportionality Measures 21

5. By-Design Principles as the Operational Foundation for Omnibus Compliance 21

3. The EU AI Act – Core Requirements for High-Risk Systems 22

1. General Introduction to the AI Act 22

2. Harmonised European Standards – Development Process and Mapping to AI Act Articles 22

3. General-Purpose AI Models and Future Standardisation Developments 25

4. Harmonised European Standards – Detailed Overview and High-Level Compliance Checklist 27

1. Detailed Overview of Candidate Harmonised European Standards 27

2. Integrated Compliance Framework for AI Systems 28

3. Integrated EU AI Act Compliance Framework – Visual Overview 29

5. Integrated Compliance Framework for MISSION 2 AI Systems – AI Act Requirements, By-Design Principles and Project-Specific Standards 31

1. EPMwDC – Espionage Prevention Monitor with Detection Capabilities 31

2. HIPSTer – Hybrid, Information, Psychological, Societal Threats Handling System 32

3. AICP-FIMI – AI Driven Cloud Platform to Counter FIMI 33

4. CTI-balanced – Cyber Threat Intelligence-based Sectorial and National Collective Cybersecurity Balanced Incentives System 34

5. OSOTS – Open Source Operational Technology Sensor 35

6. Overarching Assurance Pipeline for the MISSION 2 Portfolio 36

Conclusions and Recommendations 38

Literature 40

Executive Summary

The EU Artificial Intelligence Act (Regulation (EU) 2024/1689) and the accompanying Digital Omnibus on AI establish a risk-based regulatory framework that demands both legal compliance and disciplined engineering practice. This Compliance Report assesses the five MISSION 2 AI systems developed under the "Safe and Inclusive E-Society" programme — EPMwDC, HIPSTer, AICP-FIMI, CTI-balanced, and OSOTS — and demonstrates that they have been intentionally designed as expert and decision-support tools with limited autonomy and strong human oversight. In their current configurations, all five systems qualify as limited or minimal risk and do not trigger the full set of high-risk obligations under Annex III of the AI Act.

The report shows that proactive compliance is achieved through the systematic application of a comprehensive stack of by-design principles — risk management by design, data governance and privacy by design, cybersecurity and adversarial robustness by design, human oversight and transparency by design, accountability and traceability by design, safety and resilience by design, fairness and non-discrimination by design, ethics and democratic-impact by design, and sustainability by design. These principles are operationalised via the Integrated EU AI Act Compliance Framework presented in Figure 4.1, which aligns the AI system lifecycle (Design → Development → Evaluation → Operation → Retirement) with the relevant AI Act articles, candidate harmonised European standards (prEN 18228, prEN 18229 series, prEN 18282, prEN 18284, prEN 18286, etc.), and the concrete documentation outputs required by Annex IV.

The framework is reinforced by the Provider Quality Management System (prEN 18286) and the practical verification mapping proposed by Buscemi et al. (2026). Together, they ensure that the necessary technical documentation is generated naturally as a by-product of sound engineering rather than as a separate administrative exercise. The Digital Omnibus on AI further enhances this position by introducing targeted simplifications, proportionality measures, and strengthened sandbox opportunities that the MISSION 2 projects are well placed to exploit.

Key conclusions are that the five systems maintain a favourable risk classification when operated within their intended expert/supportive scope. Any future increase in autonomy or change in deployment context would require re-evaluation, but the embedded by-design principles already provide a robust foundation for such evolution. The report concludes with concrete, project-specific and programme-wide recommendations to maintain compliance, maximise the benefits of the Digital Omnibus, and position MISSION 2 as a model for trustworthy, human-centric AI in security-sensitive and democratic domains.

By treating compliance as an integral part of system design from the earliest TRL stages, the MISSION 2 consortium not only meets current and forthcoming regulatory requirements but also delivers innovative solutions that meaningfully contribute to a safe and inclusive European e-society.

Introduction

The European Union Artificial Intelligence Act (Regulation (EU) 2024/1689) marks a historic milestone in global technology governance. By establishing the world's first comprehensive, risk-based legal framework for artificial intelligence, the Act aims to promote the uptake of trustworthy, human-centric AI while ensuring a high level of protection for health, safety, fundamental rights, democracy, and the rule of law. Its proportionate, horizontal approach balances innovation with accountability, recognising that different AI applications carry different levels of risk and therefore require tailored obligations.

Within this evolving regulatory environment, the MISSION 2 programme – "Safe and Inclusive E-Society" – develops a portfolio of advanced yet responsible AI systems addressing critical challenges in physical security, hybrid threats, electoral integrity, collective cybersecurity, and operational technology protection. The five systems examined in this report – the Espionage Prevention Monitor with Detection Capabilities (EPMwDC), the Hybrid, Information, Psychological, Societal Threats Handling System (HIPSTer), the AI Driven Cloud Platform to Counter FIMI (AICP-FIMI), the Cyber Threat Intelligence-based Sectorial and National Collective Cybersecurity Balanced Incentives System (CTI-balanced), and the Open Source Operational Technology Sensor (OSOTS) – have been deliberately engineered as expert and decision-support tools. Characterised by limited autonomy, strong human oversight, and domain-specific custom development, they are positioned to deliver significant societal value while maintaining a favourable risk profile under the AI Act.

This compliance report provides a structured, forward-looking analysis of how the MISSION 2 AI systems align with the requirements of the EU AI Act and capitalise on the targeted simplifications introduced by the Digital Omnibus on AI. It demonstrates that proactive compliance is best achieved through the systematic application of by-design principles – risk management by design, data governance and privacy by design, cybersecurity and adversarial robustness by design, human oversight and transparency by design, accountability and traceability by design, safety and resilience by design, fairness and non-discrimination by design, ethics and democratic-impact by design, and sustainability by design – operationalised via the Integrated EU AI Act Compliance Framework presented in Chapter 4.

The report is organised as follows. Chapter 1 presents detailed descriptions of each AI system, including purpose, technologies, data handling practices, and preliminary risk classification. Chapter 2 examines the strategic implications of the Digital Omnibus on AI for the MISSION 2 portfolio. Chapter 3 outlines the core requirements of the EU AI Act for high-risk systems and introduces the candidate harmonised European standards. Chapter 4 provides a detailed overview of these standards and presents the general Integrated Compliance Framework. Chapter 5 applies this framework to each of the five MISSION 2 systems, showing how the by-design principles translate into practical, system-specific compliance strategies. The report concludes with key findings and actionable recommendations in Chapter 6, followed by the literature references.

By embedding the full stack of by-design principles from the earliest stages of development and adopting the lifecycle-oriented Integrated Compliance Framework, the MISSION 2 consortium not only ensures regulatory alignment but also creates a replicable model for trustworthy AI in security-sensitive and democratic contexts. This report therefore serves a dual purpose: as a comprehensive compliance artefact demonstrating conformity with the EU AI Act and the Digital Omnibus, and as a practical blueprint for other European R&D initiatives seeking to balance innovation, security, and societal benefit in an increasingly regulated digital landscape.

1. AI System Descriptions and Risk Overview

The EU Artificial Intelligence Act (Regulation (EU) 2024/1689, hereinafter "the AI Act"), which entered into force on 1 August 2024 and applies in stages from February 2025 onwards, establishes the first comprehensive, risk-based legal framework for the development, placing on the market, putting into service, and use of artificial intelligence in the European Union. Its primary objectives are to promote the uptake of trustworthy, human-centric AI while ensuring a high level of protection of health, safety, and fundamental rights, including democracy and the rule of law.

Definition of an AI System For the purposes of the AI Act, Article 3(1) defines an "AI system" as:

a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.

This definition is technology-neutral and focuses on the system's functional characteristics (inference, autonomy, potential adaptiveness, and environmental impact) rather than specific techniques. Commission Guidelines on the definition of an AI system (published 29 July 2025) further clarify that the concept must be interpreted flexibly to accommodate technological developments while remaining clearly bounded.

Expert Systems in the Context of the AI Act Several of the AI systems described in this report operate as expert systems or advanced decision-support systems. In the context of this document and the AI Act, an expert system refers to a narrow, domain-specific AI application that emulates the decision-making ability of a human expert in a particular field. It typically combines a knowledge base (facts, rules, and domain expertise) with an inference engine (often supported by machine-learning components for classification, anomaly detection, or pattern recognition) to provide detections, alerts, recommendations, or evidence to human operators.

Unlike general-purpose or highly autonomous AI models, expert systems are intentionally designed to assist rather than replace human judgment. They operate with meaningful human oversight (human-in-the-loop or human-on-the-loop), maintain high explainability, and limit their outputs to supportive information rather than final decisions that materially affect individuals' rights or safety-critical outcomes. This design significantly facilitates compliance with key AI Act requirements such as human oversight (Article 14), transparency, risk management (Article 9), and data governance.

All five systems presented here qualify as AI systems under Article 3(1) of the AI Act because they use machine-learning techniques for inference and generate outputs that can influence physical (e.g., security monitoring) or virtual (e.g., threat intelligence) environments. However, they are engineered as expert/supportive tools with limited autonomy and explicit human oversight. This intentional architectural choice has a direct, favourable impact on their preliminary risk classification and the applicable regulatory obligations.

Risk Classification under the EU AI Act The AI Act adopts a risk-based approach with four tiers:

- Prohibited practices (Article 5) — unacceptable risk (e.g., social scoring under Art. 5(1)(c), individual criminal-offence risk assessment or predictive policing based solely on profiling or personality traits under Art. 5(1)(d), certain manipulative or exploitative techniques).

- High-risk (Article 6 and Annex III) — systems that may adversely affect health, safety, or fundamental rights (e.g., safety components in critical infrastructure management (Annex III, point 2), certain law-enforcement or public-security applications (point 6), systems influencing democratic processes/elections (point 8)).
- Limited risk — transparency obligations only.
- Minimal risk — voluntary codes of conduct.

Risk classification is fundamentally purpose- and context-dependent. It is determined not only by technical capabilities but primarily by the intended purpose, actual deployment context, level of autonomy, presence and quality of human oversight, and the extent to which the system materially influences decisions affecting natural persons or critical infrastructure. A system that is limited/minimal risk in its core expert-system configuration can shift to high-risk (or even prohibited) if deployed with greater autonomy, integrated into automated decision-making without safeguards, used for profiling/scoring leading to detrimental treatment, or applied in sensitive domains without adequate human oversight (Art. 6(3) derogation may apply for narrow assistive tasks). Providers must maintain continuous risk management (Art. 9) and re-evaluate classification throughout the system lifecycle.

This document provides a structured description of the five artificial intelligence systems developed within the framework of ongoing R&D projects. Each system has been analysed and documented in accordance with the requirements of the AI Act, particularly the technical documentation obligations outlined in Annex IV.

The AI systems represent a diverse portfolio spanning multiple domains, including physical security, hybrid threat detection, electoral integrity protection, collective cyber threat intelligence sharing, and operational technology (OT) security in critical infrastructure. While all systems incorporate machine learning components, their scope, level of autonomy, data sources, and intended use cases vary significantly.

The descriptions follow a consistent four-part structure for each system:

1. Overall Description of the AI System (including purpose and TRL status)
2. Types of AI Models and Technologies Used
3. Data Used for Training and Operation
4. Intended Users, Use Cases and Risk Classification Considerations

A summary table is provided below to offer a comparative overview of the five AI systems. This overview highlights key differences in technological focus, data handling approaches, target environments, and TRL progression.

Table 1.1. Summary of AI Systems						
AI System	**Project Acronym**	**Primary Purpose**	**Main AI Technologies**	**Data Type**	**TRL Start → Target**	**Intended Users**
Espionage Prevention Monitor with Detection Capabilities	EPMwDC	Prevent visual leakage of sensitive information from monitors via IR noise and detect photography attempts by insiders	Computer vision, IR signal/image analysis, anomaly classification	In-house laboratory test data (images, video, IR reflections)	TRL 3 → TRL 9	Government institutions, national defence agencies, private organisations handling classified information
Hybrid, Information, Psychological, Societal Threats Handling System	HIPSTer	Detect, attribute and counter hybrid threats, disinformation, hate speech and influence operations from public sources	NLP, OSINT/SocMINT data processing, threat classification & anomaly detection	Publicly available data only (OSINT & SocMINT)	TRL 3 → TRL 8	Public security practitioners, law enforcement, government, businesses, educational institutions
AI Driven Cloud Platform to Counter FIMI	AICP-FIMI	Detect social media bots, troll farms and foreign information manipulation during elections	NLP, network analysis, anomaly detection, sentiment & content analysis	Publicly available social media data	TRL 3 → TRL 8	Election monitoring organisations, government authorities responsible for democratic processes
Cyber Threat Intelligence-based Sectorial and National Collective Cybersecurity Balanced Incentives System	CTI-balanced	Enable structured, balanced and automated cyber threat intelligence sharing at national/sector level	Threat classification, pattern recognition, collaborative data processing	Project-specific, anonymised incident and threat intelligence data	TRL 2 → TRL 8	National cybersecurity centres, sectorial organisations, critical infrastructure operators
Open Source Operational Technology Sensor	OSOTS	Real-time anomaly and cyber threat detection in Operational Technology / Industrial Control Systems	Anomaly detection ML models on network traffic	Laboratory + real-world OT/ICS network traffic data	TRL 3 → TRL 9	Critical infrastructure operators (energy, power grids, water, gas, industrial facilities)

1.1. AI System Description: Espionage Prevention Monitor with Detection Capabilities (EPMwDC)

1. Overall Description of the AI System

The Espionage Prevention Monitor with Detection Capabilities (EPMwDC) is an integrated hardware-software system designed to protect sensitive information displayed on computer monitors from unauthorised photography or video recording by a malicious operator (insider threat).

Its primary purpose is twofold: (1) to actively prevent successful capture of legible confidential data by emitting infrared (IR) noise that is invisible to the human eye but interferes with common camera sensors, and (2) to automatically detect attempts to photograph or record the screen through reflections on glass or camera lenses, trigger an alarm, capture evidence of the scene, and securely transmit it to a remote server for notification and storage.

The system is intended for use in government institutions, national defence agencies, and private organisations handling classified or high-value sensitive information. It functions as a specialised monitor that combines preventative disruption with real-time detection and response capabilities, addressing a current gap in counter-espionage and visual data leakage protection solutions. The AI components enable automated analysis of visual and infrared inputs to identify reflection patterns and support reliable decision-making regarding protective and reporting actions.

The project starts at TRL 3 (technology concept formulated, including analytical and experimental critical functions and characteristics) and aims to reach TRL 9 (actual system proven in operational environment) by the end of 2025.

2. Types of AI Models and Technologies Used

The AI system employs machine learning-based technologies, primarily in the domains of computer vision and signal/image processing, integrated into a cohesive detection and prevention pipeline.

- Computer vision and machine learning models are used in the detective module to analyse images and video streams captured by the integrated camera. These models detect and classify reflections and glares caused by camera lenses or glass surfaces, distinguishing them from false positives such as eyeglasses or surrounding objects.
- Machine learning techniques for infrared signal and image analysis support both the preventative and detective modules. They process IR noise patterns and detect IR reflections under varying angles, distances, materials, and environmental conditions.
- Additional machine learning-based predictive and classification components evaluate reflection patterns, minimise false positives and false negatives, and decide on appropriate system responses (alarm activation and evidence capture).

All AI components are custom-developed during the project. No pre-trained foundational or off-the-shelf models are used. The machine learning solutions are built and trained specifically for this application using purpose-generated test data collected through iterative laboratory experiments. The development process includes software creation, parameter fine-tuning, and continuous optimisation based on real-world testing scenarios to meet the defined performance thresholds (e.g., high text distortion rates for prevention and reliable lens detection for the detective module).

3. Data Used for Training and Operation

Training and development data: The system is developed and optimised primarily through controlled laboratory testing and iterative experimentation rather than large-scale pre-existing public datasets. Key activities include selection and analysis of hardware components (LED monitors, IR sensors, cameras), generation and testing of IR noise patterns at various frequencies and wavelengths, and capture and analysis of reflection data from different commercial lenses, glass surfaces, angles (up to 120°), distances (up to 150 cm), and environmental conditions.

Data is collected in-house during prototype development, including images and videos of reflections, glare patterns, and IR emissions under controlled and varied scenarios. False-positive sources (e.g., human eyes, glasses, ambient objects) are explicitly tested and used to refine distinction algorithms. Representativeness is pursued by covering a range of common commercial devices, materials, and workplace-like environments. The project relies on purpose-built test data generated during R&D. Continuous fine-tuning and adjustment of parameters occur throughout development based on test outcomes to meet defined success thresholds. No continuous online retraining after deployment is described; optimisation takes place during the R&D and laboratory testing phases.

Operational (inference) data: During real-world operation, the system processes live visual input from the integrated camera and IR sensors monitoring the area in front of the monitor. This includes real-time images or video of the scene and any detected reflections or glare. When a potential threat is identified, the system captures a picture of the scene and transmits it securely to the remote server. The AI processing focuses on the immediate security context. Compliance with GDPR and privacy-by-design principles is addressed separately in parallel work packages.

4. Intended Users, Use Cases and Risk Classification Considerations

Intended users are government institutions, national defence agencies, and private organisations handling classified or high-value sensitive information. Primary use cases include deployment as a specialised workstation monitor in secure environments to prevent visual data leakage through photography or video recording by malicious insiders, with automatic detection, evidence capture, and secure reporting to human security personnel.

Risk classification considerations: In its intended narrow, rule-supported, physical-environment detection and prevention role with full human oversight and no profiling or individual decision-making, the system falls under limited or minimal risk. It does not match any prohibited practices (Article 5) or high-risk categories in Annex III. Risk would remain low even if deployed at scale, provided it stays within the preventive/detection scope and does not incorporate biometric categorisation or emotion-recognition elements beyond lens/glare detection. Any future expansion (e.g., linking detections to automated access revocation without human review) would require re-assessment.

1.2. AI System Description: HIPSTer – Hybrid, Information, Psychological, Societal Threats Handling System

1. Overall Description of the AI System

The HIPSTer system is a comprehensive software platform designed to detect, attribute, and counter hybrid, information, psychological, and societal threats in the public security domain. It targets public security practitioners (law enforcement, government, military), businesses, and educational institutions.

The main purpose of the system is to automatically or semi-automatically process large volumes of publicly available data from open internet sources using OSINT (Open Source Intelligence) — the technique of collecting and analysing information from public websites, news, forums, and other open sources — and SocMINT (Social Media Intelligence) — the specialised analysis of data from public social media platforms and online communities. The system applies advanced artificial intelligence, particularly natural language processing (NLP), to identify threat indicators such as disinformation, hate speech, radical content, extremist propaganda, command-and-control communication patterns, and other hybrid threat activities.

The system aims to reduce the time between early detection of threatful activities and appropriate counteraction, while operating 24/7 with high reliability and scalability. The AI components form the core of the platform, enabling real-time or near-real-time analysis, prioritisation, and decision support for threat handling.

The project starts at TRL 3 and aims to reach TRL 8 (system complete and qualified, ready for real-world operational use) by the end of the project.

2. Types of AI Models and Technologies Used

The HIPSTer system relies on multiple machine learning and artificial intelligence technologies, primarily in the areas of natural language processing and large-scale data analysis, integrated into an automated or semi-automated threat detection and attribution pipeline. All data processed originates from publicly available sources.

- Natural language processing (NLP) and context-driven AI models are used to analyse text and content from public sources. These models support automated detection of hate speech, radical or extremist content, disinformation, misinformation, and specific communication patterns.
- Machine learning models for OSINT and SocMINT data processing enable the collection, filtering, aggregation, and analysis of large-scale publicly available data from websites and social media platforms in real time or near real time. These models help identify, prioritise, and attribute potential hybrid threats.
- Additional machine learning-based predictive and classification components are applied for threat indicator generation, anomaly detection, and decision support to distinguish genuine threats from noise and to support countermeasure recommendations.

All AI components are custom-developed during the project. The system builds upon the latest advancements in NLP, machine learning, and AI, with models being specifically trained and refined for hybrid threat detection and attribution using data collected from public sources only. Development includes iterative creation and optimisation of algorithms and models. No pre-trained foundational models from private or copyrighted sources are used; instead, the focus is on advancing the state-of-the-art through original research and tailored model development on publicly available data.

3. Data Used for Training and Operation

Training and development data: The system is developed using data collected exclusively from publicly available sources through OSINT and SocMINT techniques. Key activities include:

- Analysis of real-world case studies of hybrid threats.
- Large-scale data mining from public websites, news outlets, forums, and social media platforms.
- Collection and labelling of publicly visible text and content examples for training models on hate speech, radical content, extremist propaganda, disinformation, and other threat indicators.
- Simulation of various data collection and analysis scenarios.

No private messages, closed groups, or non-public content is used. Datasets are generated in-house during the project from open sources, ensuring they consist solely of publicly accessible information with no copyright-protected private datasets. Representativeness is pursued by covering a wide range of public online environments, including emerging platforms. False positives and edge cases are addressed during model refinement. Continuous fine-tuning and improvement of models occur throughout development and testing phases based on simulation results, laboratory testing, and field validation. Machine learning enables the system to learn from past public threat examples and improve over time.

Operational (inference) data: In operational use, the system processes live or near-real-time streams of publicly available data from open internet sources and public social media platforms using OSINT and SocMINT methods. This includes vast amounts of unstructured public text, posts, articles, and communications. The AI models analyse incoming public data continuously to detect and prioritise potential threats, generate alerts, and support attribution. When threats are identified, the system provides relevant information for counteraction. The platform is intended to operate 24/7 with high scalability and reliability in real-world public security and business environments, while respecting applicable legal frameworks for public data processing.

4. Intended Users, Use Cases and Risk Classification Considerations

Intended users are public security domain practitioners (law enforcement agencies, government officials, military), businesses, and educational institutions. Primary use cases include real-time or near-real-time monitoring of publicly available open sources and social media for early detection, attribution, and decision-support (human-led) regarding hybrid threats, disinformation, hate speech, radical content, and coordinated influence operations.

Risk classification considerations: When limited to OSINT/SocMINT-based threat intelligence, anomaly flagging, and provision of information to human operators (no automated individual decisions, no scoring leading to detrimental treatment, no predictive policing of natural persons), the system is limited risk. It processes only public data and serves an expert/supportive function.

However, risk escalates if the system is extended to:

- automated blocking, banning, or de-platforming of users/accounts based on a risk coefficient derived from social behaviour or inferred personality traits (potentially prohibited as social scoring under Article 5(1)(c));
- individual criminal-offence risk prediction or profiling (prohibited under Article 5(1)(d)); or
- direct integration into law-enforcement decision-making without meaningful human oversight.

In such scenarios the system could become prohibited or high-risk (Annex III point 6 – law enforcement). Strong governance (human oversight, audit trails, clear boundaries on automation) must be maintained to keep it within the intended limited-risk category.

1.3. AI System Description: AICP-FIMI – AI Driven Cloud Platform to Counter FIMI

1. Overall Description of the AI System

The AICP-FIMI system is an AI-driven cloud platform designed to detect and counter Foreign Information Manipulation and Interference (FIMI) during elections. Its primary purpose is to identify social media bots, troll farms, and coordinated inauthentic behaviour aimed at spreading disinformation and manipulating public opinion.

The system analyses large volumes of social media data to provide real-time alerts to authorities and election monitors, enabling timely intervention to protect the integrity of democratic elections. It combines multiple AI techniques to detect automated accounts, coordinated campaigns, linguistic patterns, network structures, and anomalous behaviour.

The project starts at TRL 3 (experimental proof of concept) and aims to reach TRL 8 (system complete and qualified for real-world operational use) by the end of the project.

2. Types of AI Models and Technologies Used

The AICP-FIMI system employs several machine learning and artificial intelligence technologies, primarily in the domains of natural language processing, network analysis, and anomaly detection, integrated into a comprehensive detection pipeline.

- Natural language processing (NLP) and text analysis models are used for sentiment analysis, content analysis, emotionality analysis, linguistic pattern recognition, and detection of repetitive or unnatural language typical of bots or coordinated campaigns.
- Machine learning models for account analysis and network analysis examine account metadata, follower/following ratios, interaction patterns, and graph-based relationships between accounts to identify inauthentic behaviour and coordinated activity.
- Anomaly detection and classification models identify unusual posting behaviour, sudden activity bursts, and deviations from normal user patterns, enabling the discovery of previously unknown bot and troll farm tactics.
- Additional predictive and multi-label classification components integrate findings from different modalities (text, network, temporal, and behavioural features) to improve overall detection accuracy and adaptability to evolving tactics.

All AI components are custom-developed during the project. Models are specifically trained and optimised for FIMI detection using project-generated datasets. The development includes iterative training, fine-tuning, and adaptation of algorithms (including embedding techniques) to handle the dynamic nature of bot and troll farm strategies. No pre-trained foundational models from external private sources are indicated as the primary basis; the focus is on creating tailored solutions through original research and hypothesis-driven experimentation.

3. Data Used for Training and Operation

Training and development data: The system is developed using datasets compiled exclusively from publicly available social media data. Key activities include:

- Collection of account information, linguistic content, network connections, temporal patterns, and interaction data from social media platforms.

- Manual annotation and labelling of accounts as genuine or inauthentic based on previously identified bot and troll farm cases.
- Creation of balanced training, validation, and test sets that include both genuine user behaviour and examples of coordinated inauthentic behaviour.
- Pre-processing and cleaning of large-scale datasets to ensure quality and representativeness.

Datasets are generated in-house during the project through systematic data gathering and annotation. The project emphasises diversity across different platforms, languages, and election contexts to improve model generalisation. Continuous refinement of models occurs based on experimental results, performance evaluation, and feedback from real-life testing scenarios. No private messages or non-public content is used.

Operational (inference) data: In real-world operation, the system processes live or near-real-time streams of publicly available data from social media platforms. This includes posts, account metadata, interaction networks, and content. The AI models continuously analyse incoming public data to detect bots, troll farms, and coordinated disinformation campaigns, generating alerts and reports for election monitors and authorities. The platform is designed for scalability and 24/7 operation in election-related monitoring contexts, while incorporating data privacy and security measures.

4. Intended Users, Use Cases and Risk Classification Considerations

Intended users are election monitoring organisations, government authorities, and institutions responsible for protecting democratic processes. Primary use cases include real-time monitoring of publicly available social media during election periods to detect bots, troll farms, coordinated inauthentic behaviour, and foreign information manipulation and interference (FIMI), generating alerts and reports for human review and timely intervention.

Risk classification considerations: In its core configuration as a detection and alerting platform supporting (not replacing) human election-integrity efforts, the system is assessed as limited risk (or at most high-risk under Annex III point 8 if it were to influence voting behaviour — which it does not; it counters manipulation). It does not perform prohibited practices.

Risk would increase if the platform were used for automated content removal, account suspension, or direct influence on electoral outcomes without human oversight, or if it incorporated real-time biometric elements. The supportive, transparency-oriented design and public-data focus keep the system comfortably outside prohibited categories and, in most deployments, outside high-risk obligations (Art. 6(3) derogation likely applies).

1.4. AI System Description: CTI-balanced – Cyber Threat Intelligence-based Sectorial and National Collective Cybersecurity Balanced Incentives System

1. Overall Description of the AI System

The CTI-balanced system is a cyber threat intelligence (CTI) management platform designed to support sectorial and national-level collective cybersecurity. Its primary purpose is to enable structured, balanced, and automated sharing of cyber threat intelligence among government agencies, critical infrastructure operators, and private entities, while addressing current gaps in incentives, quality measurement, sharing models, and integration with incident response processes.

The system aims to create practical frameworks, process models, and automation tools that facilitate effective threat intelligence collection, sharing, classification, and utilisation for improved situational awareness, incident management, and collaborative knowledge discovery. It is intended for national cybersecurity centres and sectorial organisations to implement balanced and privacy-preserving CTI practices.

The project starts at TRL 2 (technology concept formulated) and aims to reach TRL 8 (system complete and qualified for real-world operational use) by the end of the project.

2. Types of AI Models and Technologies Used

The CTI-balanced system incorporates machine learning and artificial intelligence technologies to support automation and collaborative processing of cyber threat intelligence.

- Machine learning models are used for threat classification, pattern recognition, and quality assessment of shared intelligence data.
- AI-supported data processing and analysis techniques enable collaborative processing of cybersecurity incident data in a security- and privacy-preserving environment, facilitating the discovery of new knowledge from shared threat information.
- Predictive and classification components support the automation of threat intelligence workflows, including integration with Information Security Event Management, Incident Management, and Situational Awareness services.

The AI components are custom-developed during the project to address specific scientific uncertainties in national and sectorial CTI management. Development focuses on creating tailored models and automation solutions based on research outcomes, rather than relying on pre-trained foundational models. The system combines technological automation with process frameworks, legal compliance models (including GDPR), and incentive structures.

3. Data Used for Training and Operation

Training and development data: The system is developed primarily through qualitative research and applied modelling rather than large-scale public datasets. Key activities include:

- Collection and analysis of data from ongoing client projects, interviews with national cybersecurity representatives, critical infrastructure operators, and expert working groups.
- Use of internal project data and observations from real-world incident response and threat intelligence processes.

- Development and validation of models using anonymised or structured cybersecurity incident data in privacy-preserving environments.
- Research on threat classification, incentive models, sharing parameters, and quality metrics.

Datasets are generated in-house or derived from controlled, project-specific sources with a strong emphasis on security and privacy. No large open public datasets are mentioned as the main source. Continuous refinement occurs through expert reviews, tabletop exercises, prototype testing with clients, and feedback loops. Special attention is given to GDPR compliance, privacy protection models, and legal regulatory requirements, with dedicated research conducted by MRU.

Operational (inference) data: In operational use, the system processes structured cyber threat intelligence shared voluntarily by participating organisations (government and private sector) under defined sharing models and least-impediment principles. This includes anonymised or appropriately protected incident data, threat indicators, and contextual information at strategic, operational, and tactical levels. The AI components support automated classification, quality assessment, and collaborative analysis to generate actionable insights and situational awareness. The platform is designed to operate in a secure, privacy-preserving manner compliant with GDPR and other cybersecurity regulations.

4. Intended Users, Use Cases and Risk Classification Considerations

Intended users are national cybersecurity centres, sectorial cybersecurity organisations, critical infrastructure operators, and government entities responsible for collective cyber defence. Primary use cases include structured, voluntary sharing of anonymised cyber threat intelligence, quality assessment, classification, and collaborative analysis to support (human-led) incident response and situational awareness.

Risk classification considerations: The system functions as a supportive expert tool for collective intelligence sharing and process automation in a privacy-preserving, consent-based environment. It does not involve individual profiling, decision-making affecting natural persons' rights, or safety-component roles in critical infrastructure. Therefore it qualifies as limited or minimal risk and does not fall under any Annex III high-risk category or Article 5 prohibitions. Even if scaled nationally, the balanced-incentive, human-governed, anonymised nature of the platform maintains this classification.

1.5. AI System Description: OSOTS – Open Source Operational Technology Sensor

1. Overall Description of the AI System

The Open Source Operational Technology Sensor (OSOTS) is a specialised network sensor with machine learning capabilities designed for cyber incident detection and response in Operational Technology (OT) and Industrial Control Systems (ICS).

Its primary purpose is to monitor network traffic in industrial environments (such as power plants, power grids, renewable energy farms, gas and liquid fuel facilities, and water distribution systems), detect anomalies and potential cyber threats in real time, and support security teams in maintaining the integrity and reliability of critical industrial processes. The sensor is passive, scalable, and customisable, with minimal impact on existing OT networks.

The AI components enhance traditional rule-based monitoring by enabling anomaly detection based on learned normal behaviour patterns. The project starts at TRL 3 (experimental proof of concept demonstrated in laboratory conditions) and aims to reach TRL 9 (actual system proven in operational environment through end-user testing and validation) by the end of the project.

2. Types of AI Models and Technologies Used

The OSOTS system integrates machine learning technologies focused on anomaly detection within OT/ICS network traffic.

- Machine learning models for anomaly detection are used to identify deviations from normal network behaviour, including unusual communication patterns, beaconing activity, and potential intrusions in industrial protocols.
- Feature extraction and classification techniques analyse network traffic data, topology, and protocol-specific characteristics to distinguish between legitimate operational activity and suspicious behaviour.
- The system supports both supervised and unsupervised learning approaches, with models trained to operate in real-time constraints typical of OT environments.

All AI components are custom-developed during the project. Models are specifically trained and optimised for the unique characteristics of OT/ICS networks using project-generated datasets. Development includes research on optimal statistical features and model architectures for different industrial protocols and environments. The solution combines open-source baseline components with proprietary ML modules, ensuring customisability and adaptability without relying on pre-trained foundational models from external sources.

3. Data Used for Training and Operation

Training and development data: The system is developed using datasets collected from laboratory environments and real-world OT/ICS network traffic. Key activities include:

- Collection and pre-processing of network traffic data representing normal operational behaviour across various industrial protocols.
- Creation of labelled datasets that include both normal traffic and simulated anomaly/intrusion scenarios.
- Feature selection and engineering focused on protocol-specific and network-specific characteristics.

- Iterative training, testing, and refinement of ML models using laboratory-generated and partner-provided data (e.g., from energy sector operators).

Datasets are generated in-house or obtained through controlled collaboration with industrial partners. Emphasis is placed on representativeness across different OT environments and protocols. No private or personal data processing beyond network metadata is indicated as core to the ML functions. Continuous optimisation occurs based on performance evaluation, laboratory testing, and feedback from end-user environments. Legal compliance with GDPR and cybersecurity regulations is addressed in parallel research activities.

Operational (inference) data: In real-world operation, the sensor passively captures and analyses live network traffic from OT/ICS environments. This includes communication data between industrial devices, protocol traffic, and network topology information. The machine learning models process this data in real time to detect anomalies and generate alerts without interfering with industrial processes. The system is designed for both centralised and highly distributed deployments, providing actionable intelligence to security teams while maintaining high reliability and minimal operational impact.

4. Intended Users, Use Cases and Risk Classification Considerations

Intended users are operators of critical infrastructure and industrial control systems, including energy sector organisations (power plants, grids, renewable farms), gas and fuel facilities, and water distribution companies. Primary use cases include passive real-time monitoring of OT/ICS network traffic for anomaly and cyber threat detection, generating alerts to support human security teams without interfering with industrial processes.

Risk classification considerations: Because the system is intended for use in the management and operation of critical infrastructure (energy, water, etc.), it potentially falls under Annex III, point 2 (high-risk as a safety component). However, its passive, anomaly-detection-only character, lack of direct control over physical processes, and presence of human oversight in incident response may qualify it for the Art. 6(3) derogation (“does not pose a significant risk … including by not materially influencing the outcome of decision making”). In the core intended configuration it can therefore remain limited risk provided the provider documents the narrow scope and implements appropriate safeguards.

If future iterations add autonomous control, predictive maintenance that materially affects safety, or direct integration into safety-critical actuation, it would definitively become high-risk and trigger full conformity obligations (risk management, data governance, technical documentation, etc.). It does not approach prohibited practices.

2. The Digital Omnibus on AI – Key Pillars and Strategic Implications for MISSION 2 Projects

The Digital Omnibus on AI (Proposal for a Regulation amending Regulations (EU) 2024/1689 and (EU) 2018/1139, interinstitutional file 2025/0359 (COD), Council mandate agreed on 13 March 2026) represents a pragmatic response to the first wave of implementation experience under the EU Artificial Intelligence Act. Rather than changing the Act's fundamental risk-based architecture, the Omnibus introduces targeted simplifications and clarifications designed to reduce administrative burden, improve proportionality, and accelerate responsible innovation — all while maintaining the high level of protection for health, safety, and fundamental rights.

The proposal is structured around four key pillars that directly shape the compliance landscape for the five MISSION 2 AI systems: the Espionage Prevention Monitor with Detection Capabilities (EPMwDC), the Hybrid, Information, Psychological, Societal Threats Handling System (HIPSTer), the AI Driven Cloud Platform to Counter FIMI (AICP-FIMI), the Cyber Threat Intelligence-based Sectorial and National Collective Cybersecurity Balanced Incentives System (CTI-balanced), and the Open Source Operational Technology Sensor (OSOTS). These systems, developed as custom expert/supportive tools with limited autonomy and strong human oversight, are well positioned to benefit from the Omnibus provided they continue to be engineered according to the comprehensive by-design principles set out in the Master Guide to By-Design Principles for Trustworthy and Compliant High-Stakes AI in the EU.

1. Pillar 1: Simplifications to AI Act Obligations and Conformity Procedures

The first pillar focuses on practical measures to ease the regulatory load. It streamlines conformity assessment and notified-body designation through a single application and unified assessment procedure, reduces registration obligations in the EU database for certain low-impact systems, clarifies the interplay between the AI Act and sectoral legislation listed in Annex I, and refines technical documentation requirements. These changes address the heavier-than-expected compliance burden identified during early implementation while preserving the integrity of risk management and Annex IV obligations.

For the MISSION 2 portfolio, this pillar offers tangible relief. EPMwDC, CTI-balanced, and OSOTS — which operate in controlled physical or operational environments — can leverage lighter documentation and quality management system (QMS) pathways. HIPSTer and AICP-FIMI, as software platforms processing publicly available data, similarly benefit from clearer conformity routes. By embedding accountability, traceability, and auditability by design from the earliest TRL stages, all five systems can produce the structured evidence required by Annex IV efficiently, turning simplification into a genuine competitive advantage.

2. Pillar 2: Privacy and Data Protection Enhancements

The second pillar strengthens the legal framework for bias detection and correction by extending the existing legal basis for processing special categories of personal data (under GDPR Article 9(2)(g) and equivalent rules) to a broader range of AI providers and deployers, including non-high-risk systems. Strict safeguards, necessity, and proportionality requirements remain in force.

This pillar is particularly relevant for HIPSTer and AICP-FIMI, which analyse large volumes of public text, social media, and OSINT/SocMINT data. It provides a clearer pathway for lawful bias mitigation while reinforcing the need for privacy by design and by default and data governance by design. For EPMwDC, CTI-balanced, and OSOTS, the pillar underscores the importance of purpose limitation and data minimisation in any incidental personal-data processing (e.g., camera feeds or network metadata). When these principles are applied systematically, the systems remain comfortably within their intended limited-risk classification and can make full use of the new legal basis where relevant.

3. Pillar 3: Cybersecurity and System Robustness

The third pillar places renewed emphasis on technical robustness, adversarial resilience, and cybersecurity throughout the AI lifecycle, with explicit attention to systems deployed in critical infrastructure and operational technology (OT/ICS) environments. It reinforces the expectation that AI systems must resist errors, inconsistencies, and unauthorised interference while maintaining reliable performance.

This pillar aligns closely with the operational reality of OSOTS and EPMwDC, which function in physical-security and industrial-control contexts. OSOTS, in particular, benefits from clearer expectations around passive monitoring and resilience in OT networks. CTI-balanced gains from strengthened secure-sharing controls, while HIPSTer and AICP-FIMI must ensure robustness against coordinated disinformation and adversarial inputs. By incorporating cybersecurity by design, adversarial robustness by design, and safety and resilience by design — as detailed in the Master Guide — the MISSION 2 projects not only meet these expectations but also reduce residual risk and support their preliminary limited- or minimal-risk assessments.

4. Pillar 4: Innovation Support and Proportionality Measures

The fourth pillar extends SME-specific flexibilities to small mid-cap enterprises (SMCs), replaces the mandatory AI literacy obligation with a positive encouragement by Member States and the Commission, broadens access to simplified QMS routes, strengthens regulatory sandboxes, and expands opportunities for real-world testing (including for systems covered by sectoral legislation). These measures aim to foster innovation without compromising protection.

All five MISSION 2 systems, developed in an R&D context with progressive TRL advancement, stand to gain significantly. The proportionality measures reduce compliance costs for organisations scaling from SME to SMC status. Sandbox participation and real-world testing options provide structured environments to validate HIPSTer's threat attribution, AICP-FIMI's FIMI detection, CTI-balanced's intelligence sharing, OSOTS's OT anomaly detection, and EPMwDC's reflection analysis. Early integration of human oversight by design, ethics and democratic-impact by design, risk management by design, and sustainability by design ensures that innovation remains responsible and future-proof.

5. By-Design Principles as the Operational Foundation for Omnibus Compliance

The Digital Omnibus confirms that the AI Act's simplifications and flexibilities are not automatic; they are available to providers who demonstrably implement trustworthy AI from the concept phase onward. The Master Guide to By-Design Principles for Trustworthy and Compliant High-Stakes AI in the EU supplies the exact engineering doctrine required: privacy by design, data governance by design, cybersecurity by design, risk management by design, human oversight by design, transparency and explainability by design, fairness and non-discrimination by design, accountability and traceability by design, safety and resilience by design, and sustainability by design.

For the MISSION 2 projects, embedding this full stack of principles is not an additional task — it is the most effective way to capitalise on every pillar of the Omnibus. It ensures that the systems remain expert/supportive tools with limited autonomy, preserves their favourable risk classification, produces the structured technical documentation needed for conformity, and positions the projects for seamless participation in sandboxes and real-world testing.

The chapters that follow provide a detailed mapping of each MISSION 2 system to the complete set of AI Act obligations (Chapter 3) and concrete implementation roadmaps for the by-design principles that will keep the projects compliant, innovative, and resilient under both the original AI Act and the Digital Omnibus.

3. The EU AI Act – Core Requirements for High-Risk Systems

1. General Introduction to the AI Act

The EU Artificial Intelligence Act (Regulation (EU) 2024/1689) is the world's first comprehensive, horizontal legal framework for AI. It entered into force on 1 August 2024. Originally, the majority of provisions for high-risk AI systems were scheduled to become applicable on 2 August 2026. However, the Digital Omnibus on AI has postponed these deadlines. The new application dates are 2 December 2027 for stand-alone high-risk AI systems (primarily Annex III) and 2 August 2028 for high-risk AI systems embedded as safety components in products covered by Annex I legislation.

The Act follows a risk-based approach. It prohibits a narrow set of unacceptable AI practices (Article 5), imposes specific obligations on high-risk AI systems (Chapter III, Section 2), and establishes lighter transparency obligations for certain limited-risk systems. The five MISSION 2 systems are designed as expert/supportive tools with meaningful human oversight and are therefore preliminarily assessed as limited or minimal risk in their intended configurations. However, their risk classification is purpose- and context-dependent and must be re-evaluated whenever deployment scope, autonomy level, or integration changes.

High-risk AI systems are those that (a) serve as safety components of products covered by Union harmonisation legislation listed in Annex I, or (b) fall within the specific use-cases listed in Annex III (e.g., certain law-enforcement applications, management of critical infrastructure, or systems influencing democratic processes). Providers of high-risk systems must fulfil a set of interlocking obligations:

- Risk management system (Article 9) – continuous, iterative identification, evaluation, and mitigation of known and reasonably foreseeable risks throughout the lifecycle.
- Data governance (Article 10) – high-quality, representative, relevant, and bias-managed training, validation, and testing datasets.
- Technical documentation (Article 11 and Annex IV) – comprehensive record of the system's design, development, and operation.
- Record-keeping / logging (Article 12) – automatic generation and retention of event logs to enable traceability.
- Transparency (Article 13) – clear instructions for use that enable deployers to understand capabilities, limitations, and appropriate operation.
- Human oversight (Article 14) – design features that allow natural persons to effectively oversee the system and prevent or minimise risks.
- Accuracy, robustness, and cybersecurity (Article 15) – appropriate performance levels, resilience to errors and adversarial attacks, and protection against unauthorised interference.
- Quality management system (Article 17) – documented organisational processes covering compliance, risk management, and post-market monitoring.

Conformity assessment (Article 43) and CE marking complete the obligations before market placement. Providers must also establish post-market monitoring (Article 72) and report serious incidents. Compliance with these requirements is supported by harmonised European standards (see 3.2 below), which, once cited in the Official Journal of the European Union (OJEU), confer a legal presumption of conformity.

2. Harmonised European Standards – Development Process and Mapping to AI Act Articles

Harmonised European standards (hENs) are voluntary technical specifications developed by CEN and CENELEC in response to a formal standardisation request from the European Commission. When their reference is published in the OJEU, compliance with the relevant parts of the standard creates a presumption

of conformity with the corresponding AI Act requirements. This mechanism significantly simplifies conformity assessment and reduces legal uncertainty for providers.

The development process is managed by CEN-CENELEC Joint Technical Committee 21 (JTC 21) "Artificial Intelligence". In response to the Commission's standardisation request C(2025) 3871 (which repealed the earlier request C(2023)3215), JTC 21 is preparing a horizontal portfolio of candidate hENs covering the full AI lifecycle. Each draft includes an Annex ZA that explicitly maps its technical clauses to specific AI Act articles. The process includes drafting, public enquiry (prEN stage), formal vote, and final Commission assessment before OJEU citation.

The table below provides a clear mapping of the main AI Act high-risk obligations (Articles 9–15, 17, and 43) to the current candidate harmonised standards under development.

Table 3.1. Mapping of AI Act Requirements to Candidate Harmonised European Standards

AI Act Article(s)	Requirement	Candidate Harmonised Standard(s)
Art. 9	Risk management system	prEN 18228 – AI Risk Management
Art. 10	Data quality & governance (incl. bias)	prEN 18284 – Quality & governance of datasets in AI prEN 18283 – Managing bias in AI systems
Art. 12	Record-keeping / logging	prEN ISO/IEC 24970 – AI system logging
Art. 13 & Art. 14	Transparency & Human oversight	prEN 18229-1 – AI Trustworthiness Framework – Part 1: Logging, transparency and human oversight
Art. 15 (accuracy & robustness)	Accuracy and robustness	prEN 18229-2 – AI Trustworthiness Framework – Part 2: Accuracy and robustness
Art. 15 (cybersecurity)	Cybersecurity	prEN 18282 – Cybersecurity specifications for AI systems
Art. 17	Provider quality management system	prEN 18286 – QMS for EU AI Act regulatory purposes
Art. 43	Conformity assessment	prEN 18285 – AI Conformity assessment framework

The portfolio is intentionally horizontal and lifecycle-oriented, allowing providers to address multiple requirements through a coherent set of standards rather than fragmented documents.

- **prEN 18228 (Art. 9)** establishes a structured, iterative risk management system covering identification of known and reasonably foreseeable risks, risk evaluation, mitigation controls, residual risk acceptance, and continuous monitoring.
- **prEN 18284 & prEN 18283 (Art. 10)** address data governance (dataset quality, representativeness, provenance, and management) and bias detection/mitigation practices respectively.
- **prEN ISO/IEC 24970 (Art. 12)** defines requirements for automatic logging of relevant events to ensure traceability and post-market monitoring.
- **prEN 18229-1 (Art. 13 & 14)** provides integrated technical specifications for transparency (instructions for use, explainability) and human oversight (design features enabling effective operator intervention).
- **prEN 18229-2 (Art. 15 accuracy & robustness)** covers performance metrics, robustness testing, error handling, and resilience to distribution shifts and adversarial attacks.
- **prEN 18282 (Art. 15 cybersecurity)** specifies cybersecurity controls tailored to AI systems, including protection against data poisoning, model evasion, and unauthorised access.
- **prEN 18286 (Art. 17)** outlines the organisational quality management system that providers must implement to ensure ongoing compliance.

- **prEN 18285 (Art. 43)** describes the conformity assessment procedures, including self-assessment and third-party (notified body) routes.

These candidate standards, once finalised and cited in the OJEU, will offer providers of high-risk AI systems (and those MISSION 2 systems that may fall into high-risk categories depending on final deployment) a practical, technically detailed pathway to demonstrate conformity. Until citation occurs, providers may still demonstrate compliance using state-of-the-art methods and comprehensive technical documentation as required by Annex IV.

The MISSION 2 systems, being custom-developed expert tools with strong human oversight, are positioned to leverage these standards efficiently. Early alignment with the forthcoming hENs (particularly risk management, data governance, cybersecurity, and human oversight) will further strengthen their compliance posture and support their intended limited- or minimal-risk classification.

Recent scholarship provides a practical bridge between the legal obligations of the AI Act and the technical verification activities needed for conformity assessment. Buscemi et al. (2026) present a structured operational mapping that decomposes the high-level requirements of the Regulation (human oversight, technical robustness and safety, data governance, transparency, diversity and non-discrimination, societal and environmental well-being, accountability, quality management, risk management, technical documentation, and record-keeping) into concrete, implementable verification activities. These activities are organised along two dimensions: type of verification (controls versus empirical testing) and lifecycle target (data, models, processes, final product). The framework directly supports the conformity assessment procedures under Articles 43 and 44 and Annexes VI–VII and serves as a reusable reference for providers, notified bodies, and market-surveillance authorities when applying the candidate harmonised standards under Standardisation Request M/613. For the MISSION 2 systems, which are currently positioned as limited- or minimal-risk expert tools, this verification mapping offers a forward-looking assurance template that can be adopted if any future evolution increases autonomy or triggers high-risk classification.

3. General-Purpose AI Models and Future Standardisation Developments

While the current portfolio of candidate harmonised European standards (detailed in Section 3.2) is specifically designed to support compliance with the high-risk obligations in Chapter III, Section 2 of the AI Act, the Regulation also contains dedicated rules for general-purpose AI (GPAI) models in Chapter V (Articles 51–55 and Annexes XI–XII). These rules apply to foundational models that can be adapted for a wide range of downstream tasks and impose transparency, documentation, and (for models presenting systemic risk) additional risk-assessment and mitigation obligations.

At present, the five MISSION 2 AI systems (EPMwDC, HIPSTer, AICP-FIMI, CTI-balanced, and OSOTS) are custom-developed expert/supportive tools that rely on domain-specific, purpose-built machine-learning components rather than general-purpose AI models. They do not incorporate GPAI models and are therefore not subject to Chapter V obligations. Their expert/supportive character, limited autonomy, and strong human oversight further reinforce their preliminary limited- or minimal-risk classification under the AI Act.

However, future evolution of these systems — particularly any extension toward greater autonomy, multi-agent orchestration, or agentic workflows — could involve the integration or adaptation of GPAI models. In such scenarios, the systems would need to address the specific transparency, documentation, and systemic-risk requirements applicable to GPAI.

To support implementation of these Chapter V obligations, the AI Act foresees a voluntary General-Purpose AI Code of Practice (third draft published in March 2025). The Code provides non-binding recommendations and practical measures for providers of GPAI models, focusing on transparency and copyright compliance (Article 53(1)(a)–(c)), as well as risk assessment, technical mitigation, and governance for models presenting systemic risk (Article 55). It is not a regulation and does not confer a presumption of conformity, but it serves as an important reference point for good practice and is expected to inform future standardisation work.

Looking ahead, the European Commission is preparing additional standardisation requests under Article 40 of the AI Act. As highlighted in the AI Board's report on possible elements of forthcoming additional standardisation requests, these future requests are likely to address:

- horizontal GPAI needs, including multi-agent and orchestration safety, evaluation methods, and governance of agentic systems;
- interfaces between GPAI models and high-risk AI systems (e.g., when GPAI components are embedded in Annex III use cases);
- alignment with existing ICT, cybersecurity, and information-security standards to avoid fragmentation; and
- sustainable AI aspects, particularly resource and energy efficiency measurement, reporting, and lifecycle considerations (Article 40(2)).

These additional deliverables will complement the existing high-risk horizontal portfolio and provide more targeted technical specifications for GPAI-related compliance.

For the MISSION 2 portfolio, the current absence of GPAI components means that Chapter V and the GPAI Code of Practice have no immediate applicability. Nevertheless, the projects are well positioned to benefit from future standardisation outputs. By continuing to embed the by-design principles from the Master Guide (particularly risk management by design, human oversight by design, transparency/explainability by design, and cybersecurity by design), the systems will remain adaptable to any future requirements that may arise if autonomy levels increase or GPAI elements are introduced. Early monitoring of the GPAI Code of Practice and forthcoming additional standardisation requests will ensure a smooth transition should the projects evolve in this direction.

In summary, the current harmonised standards provide the immediate technical backbone for high-risk obligations, while the GPAI Code of Practice and anticipated additional standardisation requests under Article 40 represent the evolving framework for general-purpose and more autonomous AI applications. The

MISSION 2 systems, in their present expert/supportive configuration, fall comfortably outside these additional layers but are engineered to incorporate them seamlessly if and when required.

Recent scholarship further highlights the regulatory implications of moving toward greater autonomy. Nannini et al. (2026) analyse AI agents—systems that autonomously plan, invoke external tools, and execute multi-step action chains with reduced human involvement. Although the five MISSION 2 AI systems currently rely on purpose-built, domain-specific components rather than general-purpose AI models or agentic workflows, any future extension toward autonomous decision-making would likely engage GPAI components and trigger additional obligations under Chapter V of the AI Act. Nannini et al. identify agent-specific compliance challenges that are only partially addressed by the existing harmonised standards portfolio: privilege minimisation outside the generative model (prEN 18282), oversight evasion risks in reinforcement-learning loops (prEN 18229-1), transparency across multi-party action chains, and the boundary between anticipated adaptive behaviour and substantial modification under Article 3(23) (prEN 18228). Their analysis also underscores the interplay with the voluntary General-Purpose AI Code of Practice and the forthcoming additional standardisation requests under Article 40. The MISSION 2 projects' continued emphasis on strong human oversight, limited autonomy, and by-design principles therefore provides a solid foundation for maintaining limited- or minimal-risk classification even under such extensions.

4. Harmonised European Standards – Detailed Overview and High-Level Compliance Checklist

The candidate harmonised European standards developed under CEN-CENELEC Joint Technical Committee 21 provide practical technical specifications that support providers in demonstrating conformity with the obligations for high-risk AI systems under the EU AI Act. When their references are published in the Official Journal of the European Union, compliance with the relevant parts of these standards creates a legal presumption of conformity. The portfolio is intentionally horizontal and lifecycle-oriented, enabling organisations to address multiple AI Act requirements through a coherent set of documents rather than fragmented approaches.

Recent scholarship further strengthens this framework by providing a structured operational mapping from high-level AI Act requirements to concrete, verifiable assessment activities. Buscemi et al. (2026) decompose the Regulation's obligations into implementable verification activities organised along two dimensions: type of verification (controls versus empirical testing) and lifecycle target (data, models, processes, final product). This mapping serves as a practical bridge between legal obligations and technical assurance practices, complementing the harmonised standards and supporting consistent compliance verification.

1. Detailed Overview of Candidate Harmonised European Standards

prEN 18228 – AI Risk Management (Art. 9) This standard establishes a structured, iterative risk management system covering the identification of known and reasonably foreseeable risks, risk evaluation, mitigation controls, residual risk acceptance, and continuous monitoring throughout the AI system lifecycle. To conform to this standard, organisations typically prepare a risk management methodology that defines risk assessment criteria, scales, and acceptance thresholds; detailed risk assessment records that systematically identify foreseeable risks to health, safety, and fundamental rights; a risk treatment plan that documents mitigation measures, responsible parties, and timelines; residual risk acceptance documentation that justifies which risks remain after treatment; and procedures for ongoing risk monitoring and periodic re-evaluation.

prEN 18284 – Quality & governance of datasets in AI and prEN 18283 – Managing bias in AI systems (Art. 10) These standards address data quality and governance, including bias management. prEN 18284 focuses on the quality and governance of datasets used for training, validation and testing. prEN 18283 provides concepts, measures and requirements for identifying, assessing and mitigating unwanted bias across data, model, system and socio-technical levels. To conform to these standards, organisations typically prepare dataset specification documents, data quality criteria registers, bias profile documentation, procedures for identification of relevant groups and bias-related hazards, bias metrics registers, bias estimation and evaluation records, mitigation measure registers, and ongoing bias monitoring procedures.

prEN ISO/IEC 24970 – AI system logging (Art. 12) This standard defines requirements for automatic record-keeping and logging of relevant events to ensure traceability, support post-market monitoring and enable investigation of incidents or substantial modifications. To conform to this standard, organisations typically prepare procedures for identification of relevant events, logging system design and specification documents, event log templates, storage and retention policies, technical documentation of logging capabilities, and procedures for post-market log review and corrective actions.

prEN 18229-1 – AI Trustworthiness Framework – Part 1: Logging, transparency and human oversight (Arts. 13 & 14) This part of the trustworthiness framework provides guidance on implementing logging capabilities, transparency mechanisms and effective human oversight arrangements. To conform to this standard, organisations typically prepare logging system design documents, transparency assessment reports, instructions for use, human oversight procedures, designated roles and responsibilities documentation, human intervention logs, training and awareness plans, verification and validation reports, and post-market monitoring procedures.

prEN 18229-2 – AI Trustworthiness Framework – Part 2: Accuracy and robustness (Art. 15 – accuracy & robustness) This standard addresses functional correctness (accuracy) and robustness of AI systems, covering performance evaluation methods, testing under varying conditions, and measures to ensure reliable operation throughout the lifecycle. To conform to this standard, organisations typically prepare accuracy assessment procedures, robustness design and monitoring documentation, test data management procedures, group-specific analysis reports, technical documentation of accuracy and robustness measures, instructions for use, training plans, and ongoing monitoring and corrective action records.

prEN 18282 – Cybersecurity specifications for AI systems (Art. 15 – cybersecurity) This standard specifies cybersecurity requirements tailored to AI systems, focusing on the identification of AI-specific threats and vulnerabilities and the implementation of appropriate technical and organisational controls. To conform to this standard, organisations typically prepare a cybersecurity framework document, procedures for identification of AI-specific vulnerabilities and threats, cybersecurity risk determination records, measures implementation registers, verification and testing procedures, technical documentation of the cybersecurity framework, and ongoing monitoring and reassessment procedures.

prEN 18286 – QMS for EU AI Act regulatory purposes (Art. 17) This standard defines the requirements for a quality management system specifically designed to support compliance with the EU AI Act, covering organisational processes for design, development, verification, post-market monitoring, incident reporting and continual improvement. To conform to this standard, organisations typically prepare a quality policy, roles and responsibilities documentation, procedures for regulatory requirements identification and QMS scope definition, quality objectives and planning records, resource and competence management procedures, product realisation and verification procedures, post-market monitoring and incident reporting procedures, performance evaluation processes, and technical documentation of the QMS.

prEN 18285 – AI Conformity assessment framework (Art. 43) This standard provides the overall framework and procedures for conformity assessment of AI systems, including internal control and third-party assessment routes, documentation requirements and CE marking processes. To conform to this standard, organisations typically prepare conformity assessment plans, technical documentation packages, quality management system records, and evidence demonstrating fulfilment of all applicable AI Act requirements.

2. Integrated Compliance Framework for AI Systems

To translate the high-level obligations of the EU AI Act and the supporting candidate harmonised European standards into operational reality, providers should adopt an integrated compliance framework grounded in the full set of by-design principles. This framework organises all compliance activities around the AI system lifecycle, maps each phase to the relevant AI Act requirements and harmonised standards, and ensures that the necessary technical documentation and evidence are generated as a natural outcome of sound engineering.

The following table summarises how the AI system lifecycle aligns with the core AI Act requirements, the candidate harmonised European standards, and the main types of documentation organisations should produce. This structure is further supported by operational verification mappings such as those proposed by Buscemi et al. (2026), which decompose requirements into concrete controls and testing activities across the lifecycle targets of data, models, processes, and final product.

Table 4.1. Mapping of AI System Lifecycle Phases to AI Act Requirements, Harmonised Standards and Key Documentation

Lifecycle Phase	Key AI Act Requirements	Supporting Harmonised Standards	Typical Documentation / Evidence Produced
1. Scoping and Intended Purpose	Art. 6 (classification), Art. 8 (intended purpose)	prEN 18285, prEN 18286	Classification memo, intended-purpose statement, high-level risk register

2. Data Acquisition and Governance	Art. 10 (data quality & governance)	prEN 18284, prEN 18283	Dataset specification, data quality criteria register, bias profile, data provenance records
3. Model Development and Training	Art. 9 (risk management), Art. 15 (accuracy & robustness)	prEN 18228, prEN 18229-2	Model architecture description, training process records, robustness test results, threat model
4. Verification, Validation and Testing	Art. 15 (accuracy, robustness, cybersecurity)	prEN 18229-2, prEN 18282	Verification and validation reports, adversarial test results, performance metrics, subgroup analysis
5. Integration, Human Oversight and Deployment	Art. 13 (transparency), Art. 14 (human oversight)	prEN 18229-1	Instructions for use, human oversight procedure, oversight roles and responsibilities, logging system design
6. Operation, Monitoring and Post-Market	Art. 12 (logging), Art. 72 (post-market monitoring)	prEN ISO/IEC 24970, prEN 18286	Event logs, post-market monitoring plan, performance monitoring records
7. Change Management and Decommissioning	Art. 9, Art. 11, Art. 17	prEN 18228, prEN 18286	Change assessment records, version history, decommissioning plan

3. Integrated EU AI Act Compliance Framework – Visual Overview

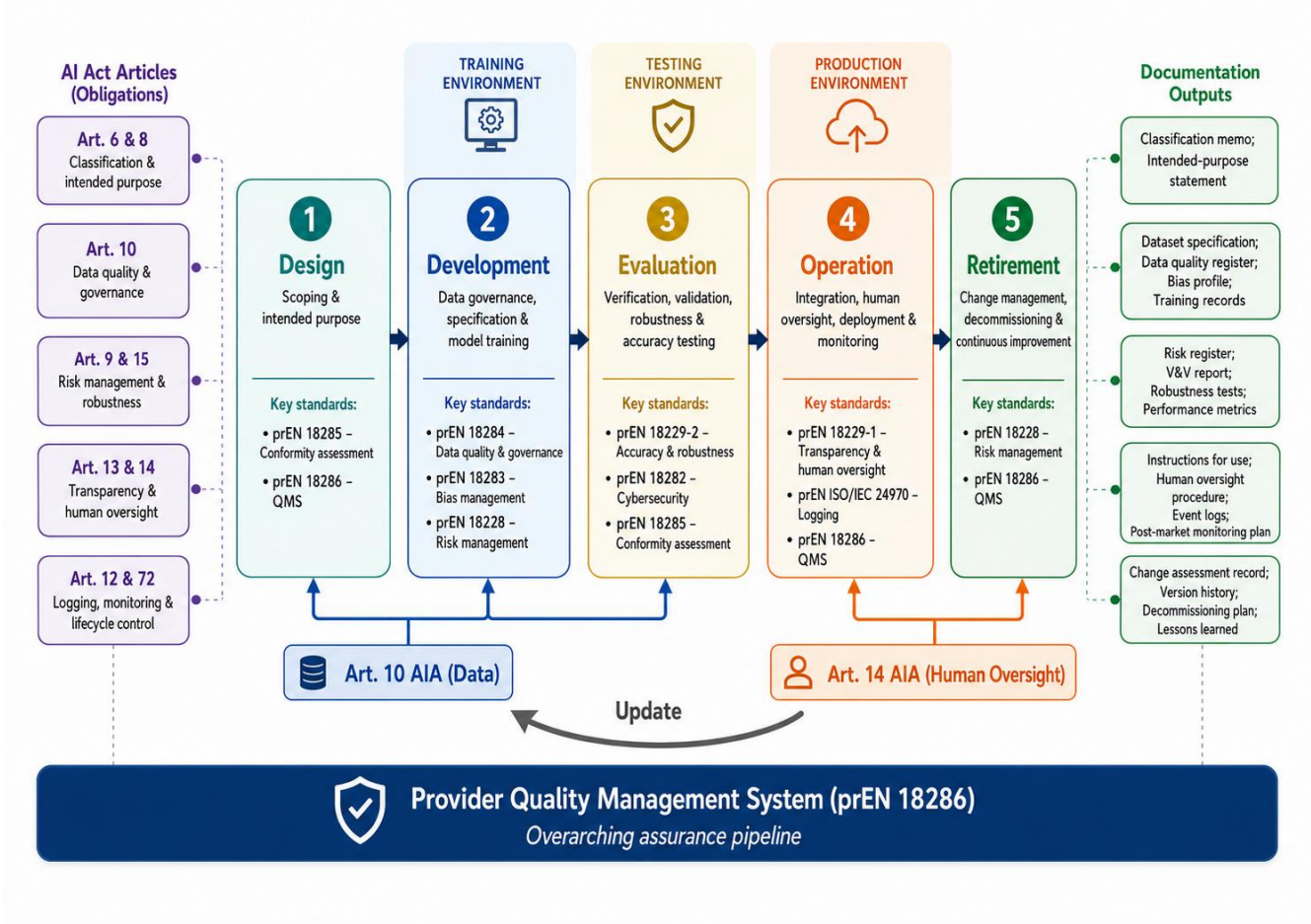

Figure 4.1 Integrated EU AI Act Compliance Framework

Figure 4.1 presents the Integrated EU AI Act Compliance Framework. It provides a clear, operational roadmap that aligns the AI system lifecycle with the core obligations of the EU AI Act, the candidate

harmonised European standards, and the concrete documentation outputs required for Annex IV technical documentation.

The diagram follows a horizontal five-phase AI lifecycle flow, supported by three distinct coloured environments (Training, Testing, and Production) that reflect real-world development and deployment realities. On the left side, purple boxes list the key AI Act articles and obligations that apply at each phase. Under each lifecycle phase, the corresponding candidate harmonised European standards are shown in blue. On the right side, green boxes detail the primary documentation outputs that must be produced to demonstrate conformity. Two critical horizontal connections at the bottom highlight Art. 10 (Data) and Art. 14 (Human Oversight), with a central "Update" feedback loop illustrating the iterative nature of compliance. The entire framework is governed by the Provider Quality Management System (prEN 18286), shown as the wide bottom bar and described as the "overarching assurance pipeline".

This visual directly embeds the by-design principles that transform regulatory obligations into disciplined engineering practice. Risk management by design runs through every phase via prEN 18228, ensuring that known and reasonably foreseeable risks are identified, assessed, mitigated, and continuously monitored. Data governance by design and privacy by design are anchored in Phase 2 (Development) through prEN 18284 and prEN 18283, enforcing purpose limitation, data minimisation, representativeness, bias mitigation, and privacy-enhancing techniques so that personal data is processed only when strictly necessary. Cybersecurity by design and adversarial robustness by design are operationalised in Phases 3 and 4 via prEN 18282 and prEN 18229-2, requiring threat modelling, secure development practices, adversarial testing, and resilient architectures that protect against poisoning, evasion, and model degradation. Human oversight by design and transparency by design are central to Phase 4 (Operation) through prEN 18229-1, mandating clear instructions for use, explainability mechanisms, operator training, override capabilities, and uncertainty signalling to prevent automation bias and ensure meaningful human control. Accountability by design is realised through logging (prEN ISO/IEC 24970) and version-controlled records across all phases, while safety, resilience, and continuity by design ensure safe degradation modes and business continuity in critical environments.

By following these by-design principles, organisations do not treat compliance as an after-the-fact documentation exercise. Instead, the required technical documentation (classification memo, dataset specification, risk register, instructions for use, post-market monitoring plan, etc.) is generated naturally as a by-product of sound engineering. The Provider Quality Management System (prEN 18286) acts as the unifying governance layer that integrates all principles, controls evidence production, and supports continuous improvement and post-market monitoring.

This framework demonstrates that EU AI Act conformity is achieved not through isolated checklists but through a coherent, lifecycle-wide engineering discipline. When the by-design principles are systematically applied, high-risk AI systems become both more trustworthy and more demonstrably compliant.

5. Integrated Compliance Framework for MISSION 2 AI Systems – AI Act Requirements, By-Design Principles and Project-Specific Standards

Figure 4.1 provides the general Integrated EU AI Act Compliance Framework for AI systems. It links the main AI Act obligations to the AI system lifecycle and shows how compliance should be generated through five connected stages: Design, Development, Evaluation, Operation, and Retirement. The framework also shows that each stage should produce specific documentation outputs, while the whole process is governed by the Provider Quality Management System under prEN 18286.

The purpose of this chapter is to adapt that general framework to the five MISSION 2 AI systems. This adaptation is necessary because the systems have different technical architectures, data environments, operational contexts, and risk profiles. EPMwDC operates in a controlled physical-security environment; HIPSTer and AICP-FIMI analyse public information spaces; CTI-balanced supports collective cyber threat intelligence sharing; and OSOTS monitors operational technology and industrial control systems. Therefore, the same framework cannot be applied mechanically. It must be interpreted in relation to the intended purpose, data sources, autonomy level, human oversight model, and foreseeable misuse of each system.

At the same time, the five systems share a common compliance logic. They are designed as expert or decision-support tools rather than autonomous decision-making systems. This is the central assumption that supports their favourable risk classification under the AI Act. Figure 4.1 therefore functions as a practical control model: it ensures that the intended purpose is fixed at the design stage, data governance is embedded during development, robustness is tested before deployment, human oversight is maintained during operation, and changes are reassessed during retirement or major updates.

In this chapter, each system is analysed through the same lifecycle structure. For each project, the framework clarifies how AI Act obligations, by-design principles, harmonised standards, documentation outputs, and risk controls should be adapted in practice.

1. EPMwDC – Espionage Prevention Monitor with Detection Capabilities

For EPMwDC, the framework in Figure 4.1 should be applied as a controlled-environment security assurance model. The system has a narrow and clearly defined purpose: to prevent visual leakage of sensitive information from computer monitors and to detect possible attempts to photograph or record the screen. This narrow purpose is important because it keeps the system outside broader categories such as biometric identification, behavioural profiling, or automated enforcement.

At the Design stage, the main task is to document the intended purpose and classification boundary. The classification memo should define EPMwDC as a preventive and detective physical-security tool. It should explain that the AI component analyses infrared signals, lens reflections, glare patterns, and related visual indicators in order to detect possible recording attempts. The same document should also state the system's limits: it does not identify individuals, does not infer intention or emotional state, does not create a personal risk score, and does not automatically impose sanctions. This directly corresponds to the first block in Figure 4.1, where Articles 6 and 8 require classification and intended-purpose definition.

At the Development stage, the framework should focus on Article 10 data governance and Article 9 risk management. EPMwDC relies mainly on laboratory-generated visual and infrared data, including reflections from lenses, glass surfaces, eyeglasses, and other objects. The dataset specification should therefore describe the test environments, angles, distances, lighting conditions, materials, sensor settings, and false-positive sources used during model development. Because camera input may incidentally capture people or workplace scenes, privacy by design should be implemented through data minimisation, short retention periods, secure evidence transfer, and strict access control. These measures connect the project to the data-governance and privacy functions shown in Figure 4.1.

At the Evaluation stage, the emphasis should be on robustness and accuracy. EPMwDC must be tested not only under ideal laboratory conditions but also under realistic workplace conditions. The verification and validation report should show how the system performs with different monitor types, camera models,

reflection angles, lighting levels, glasses, glass surfaces, and deliberate spoofing attempts. This stage corresponds to the evaluation block in Figure 4.1 and should rely mainly on prEN 18229-2 for accuracy and robustness, prEN 18282 for cybersecurity, and prEN 18228 for risk management.

At the Operation stage, human oversight becomes the key safeguard. Alerts generated by EPMwDC should be treated as security indicators requiring human review, not as final determinations of misconduct. Instructions for use should explain what an alert means, what uncertainty remains, and what the responsible operator must check before any follow-up action. Event logs should capture the time of the alert, the technical trigger, the system configuration, the evidence generated, and the human response. This connects directly with the Article 14 human oversight function and Article 12 logging function shown in Figure 4.1.

At the Retirement or change-management stage, the main issue is scope control. If EPMwDC is later connected to automated access revocation, disciplinary procedures, biometric identification, or broader workplace surveillance, the risk classification would need to be reassessed. The decommissioning and change-assessment record should therefore document whether the system remains within its original preventive and detective scope.

The expected documentation outputs for EPMwDC are the classification memo, intended-purpose statement, dataset specification, privacy safeguards, risk register, robustness testing report, instructions for use, human-oversight procedure, event logs, and change-assessment record. In its current configuration, the system can remain limited or minimal risk, provided that human review is preserved and the system is not expanded into biometric or automated enforcement functions.

2. HIPSTer – Hybrid, Information, Psychological, Societal Threats Handling System

HIPSTer requires a different adaptation of Figure 4.1 because it operates in the public information environment. Its purpose is to support detection, attribution, and analysis of hybrid threats, disinformation, hate speech, radical content, and coordinated influence operations based on OSINT and SocMINT data. The main compliance challenge is not only technical accuracy but also the protection of fundamental rights, democratic participation, freedom of expression, and non-discrimination.

At the Design stage, the framework should clearly define HIPSTer as an expert-support and intelligence-analysis platform. The intended-purpose statement should explain that the system assists human analysts by identifying patterns, indicators, and anomalies in publicly available information. It should also clarify that HIPSTer is not an automated law-enforcement decision system, not a predictive-policing tool, and not an automated content-removal mechanism. This distinction is essential for maintaining the limited-risk classification and avoiding prohibited or high-risk uses.

At the Development stage, Article 10 data governance becomes central. Public availability of data does not eliminate the need for responsible data governance. The dataset specification should document the public sources used, inclusion and exclusion criteria, language coverage, annotation methods, labelling rules, and known limitations. Since the system may analyse politically sensitive or socially contested content, bias management should be embedded into development rather than treated as a later compliance add-on. The bias profile should examine whether the system performs differently across languages, topics, communities, political contexts, or minority groups. This corresponds directly to the development block in Figure 4.1, especially prEN 18284 on data quality and governance and prEN 18283 on bias management.

At the Evaluation stage, HIPSTer should be tested against both technical and socio-technical risks. Standard accuracy metrics are not sufficient. The verification process should examine false positives in lawful political speech, satire, sarcasm, activist communication, minority-language content, and rapidly changing hybrid-threat narratives. Evaluation should also include robustness against adversarial information operations, coordinated manipulation, and concept drift. In this way, the evaluation stage applies Article 15 robustness requirements together with the democratic-impact and fairness-by-design principles.

At the Operation stage, the system should provide explainable alerts and evidence packages rather than unexplained risk labels. Human analysts should be able to see why content, a narrative, or a network cluster was flagged, which indicators contributed to the alert, and what confidence level is attached to the output. The system should also support analyst override and correction. Event logs should record the input source,

model version, alert rationale, analyst decision, and any escalation. This operational model follows Figure 4.1 by connecting transparency, human oversight, logging, and post-market monitoring.

At the Retirement or change-management stage, HIPSTer should be reassessed whenever its outputs are connected to stronger operational consequences. If the system is later used to support automated blocking, user scoring, account banning, criminal-risk prediction, or direct law-enforcement action without meaningful human oversight, its classification could change substantially. The change-assessment record should therefore track whether the system remains an advisory intelligence-support tool.

The expected documentation outputs for HIPSTer are the classification memo, intended-purpose statement, OSINT/SocMINT dataset specification, data quality register, bias profile, risk register, verification and validation report, instructions for use, analyst oversight procedure, event logs, and post-market monitoring plan. HIPSTer remains limited risk when it supports human-led threat analysis, but its compliance position depends on keeping clear boundaries around automation and individual-level consequences.

3. AICP-FIMI – AI Driven Cloud Platform to Counter FIMI

AICP-FIMI should be adapted to Figure 4.1 as a democratic-integrity assurance system. Its objective is to detect bots, troll farms, coordinated inauthentic behaviour, and foreign information manipulation and interference during election periods. Because the system operates in the electoral context, its compliance framework must give particular attention to democratic impact, transparency, human oversight, and proportionality.

At the Design stage, the classification memo should define AICP-FIMI as a detection, alerting, and situational-awareness platform for election monitors and competent authorities. Its intended purpose is to help identify coordinated manipulation, not to determine the legitimacy of political speech or directly influence voters. This distinction is essential. The system should not be designed as an automated content moderation tool, account-suspension tool, or public attribution mechanism. In the terminology of Figure 4.1, the design stage must fix the boundary between supportive analysis and automated intervention in democratic processes.

At the Development stage, Article 10 data governance should focus on the construction of representative and balanced datasets from publicly available social media data. The dataset specification should describe account metadata, linguistic content, network features, temporal patterns, annotation criteria, and the distinction between genuine political participation and coordinated inauthentic behaviour. This distinction is particularly important during elections because legitimate political activity may also be intense, emotional, coordinated, and repetitive. The framework therefore requires careful labelling logic and bias analysis across platforms, languages, political contexts, and election scenarios.

At the Evaluation stage, AICP-FIMI should be tested under realistic election-period conditions. Evaluation should not only measure whether known bots or troll networks are detected. It should also assess whether the system avoids over-alerting during legitimate spikes in political communication. Robustness testing should include adversarial adaptation, changing bot strategies, multilingual manipulation, coordinated posting, artificial amplification, and sudden changes in network behaviour. This links the project to the evaluation block in Figure 4.1, especially prEN 18229-2 for accuracy and robustness and prEN 18282 for cybersecurity.

At the Operation stage, the system should generate contextualised alerts for human review. Alerts should indicate whether suspicious behaviour is based on linguistic similarity, temporal coordination, account metadata, network structure, or a combination of indicators. Election monitors should receive sufficient explanation to understand both the signal and its uncertainty. Operational use should remain human-led: the system may prioritise cases and support situational awareness, but it should not automatically remove content, suspend accounts, issue public accusations, or trigger enforcement action. This reflects the Article 14 human oversight and Article 13 transparency components shown in Figure 4.1.

At the Retirement or change-management stage, AICP-FIMI should be reassessed before every new election deployment or major platform integration. Election environments differ across countries, languages, platforms, and political contexts. A model that performs well in one election may not be appropriate in

another without recalibration. The change-assessment record should therefore document new data sources, model updates, changed monitoring objectives, and any new operational consequences attached to system outputs.

The expected documentation outputs for AICP-FIMI are the classification memo, intended-purpose statement, social media dataset specification, annotation methodology, bias profile, risk register, robustness and adversarial testing reports, instructions for use, human-review procedure, event logs, post-market monitoring plan, and election-specific change-assessment records. The system can remain limited risk when it supports human-led detection and monitoring. However, if it is used to materially influence electoral processes, automate de-platforming, or directly affect political participation, its classification would need to be reconsidered.

4. CTI-balanced – Cyber Threat Intelligence-based Sectorial and National Collective Cybersecurity Balanced Incentives System

CTI-balanced should be adapted to Figure 4.1 as a governance-oriented cybersecurity intelligence platform. Its purpose is to support structured, balanced, and privacy-preserving cyber threat intelligence sharing between national cybersecurity centres, sectorial organisations, critical infrastructure operators, and other participating entities. The main compliance question is how to ensure that automation improves collective cybersecurity without undermining confidentiality, trust, data quality, or accountability.

At the Design stage, the intended-purpose statement should define CTI-balanced as a collaborative decision-support and intelligence-sharing system. It should clarify that the platform supports classification, quality assessment, pattern recognition, and situational awareness, but does not replace human responsibility for incident response or national-level cybersecurity decisions. The risk register should consider not only AI model risks but also ecosystem-level risks such as false intelligence, intelligence poisoning, incentive gaming, unauthorised disclosure, re-identification of anonymised incident data, and over-reliance on automated scores.

At the Development stage, Article 10 data governance should be adapted to the specific nature of cyber threat intelligence. CTI data may include incident descriptions, indicators of compromise, organisational context, technical artefacts, and sometimes sensitive operational information. The dataset specification should therefore define what types of intelligence may be shared, how they are anonymised or pseudonymised, who may access them, and how quality and provenance are recorded. The framework should also distinguish between strategic, operational, and tactical intelligence because each level may require different protection and sharing rules.

At the Evaluation stage, the system should be tested for both cybersecurity resilience and intelligence quality. Technical testing should examine unauthorised access, manipulation of shared indicators, poisoning of intelligence feeds, and confidentiality breaches. At the same time, validation should assess whether automated classification and quality scoring genuinely help human users evaluate relevance, reliability, and urgency. This connects CTI-balanced to prEN 18282 for cybersecurity, prEN 18284 for data governance, prEN 18228 for risk management, and prEN 18229-1 for human oversight.

At the Operation stage, human governance is central. Participating organisations must be able to understand how intelligence is classified, how confidence is assessed, and how information is shared or restricted. Logs should record who submitted intelligence, how it was classified, who accessed it, whether it was corrected or withdrawn, and how it contributed to later decisions. This reflects the logging and post-market monitoring functions in Figure 4.1. The platform should support trust between participants by making accountability visible and by ensuring that no automated classification becomes an unquestioned decision.

At the Retirement or change-management stage, CTI-balanced should maintain clear procedures for removing obsolete indicators, withdrawing incorrect intelligence, updating sharing rules, and preserving necessary audit records. Decommissioning is especially important in cybersecurity because outdated intelligence, old access rights, and unmaintained integrations can become vulnerabilities.

The expected documentation outputs for CTI-balanced are the classification memo, intended-purpose statement, data-sharing policy, dataset and provenance specification, access-control model, cybersecurity framework, risk register, quality assessment methodology, instructions for use, event logs, incident-handling procedure, and change-assessment record. In its current form, CTI-balanced remains limited or minimal risk because it supports anonymised, consent-based, human-governed cyber threat intelligence sharing. Its classification would need reassessment only if it were extended toward automated enforcement, individual profiling, or direct safety-critical decision-making.

5. OSOTS – Open Source Operational Technology Sensor

OSOTS is the system that requires the most careful adaptation of Figure 4.1 because it is intended for operational technology and industrial control system environments, including energy, water, gas, and other critical infrastructure contexts. Even though the current system is passive and advisory, its deployment environment is high-stakes. Therefore, the framework must show clearly why OSOTS is not a safety component in its current configuration and how the passive monitoring boundary is preserved.

At the Design stage, the classification memo should define OSOTS as a passive anomaly-detection sensor. It monitors OT/ICS network traffic, identifies deviations from expected behaviour, and generates alerts for human security teams. It does not autonomously control industrial processes, change operational parameters, shut down equipment, or trigger safety actions. This intended-purpose statement is central to the risk classification. It supports the argument that the system may remain limited risk, or may qualify for an Article 6(3)-type reasoning where it does not materially influence decision-making outcomes, provided the passive and advisory scope is maintained.

At the Development stage, Article 10 data governance should be adapted to OT/ICS network data. The dataset specification should document industrial protocols, traffic features, topology characteristics, normal operating patterns, simulated attack scenarios, and partner-provided data sources. Because OT environments are highly specific, the framework should avoid assuming that one generic baseline will work everywhere. Data governance should therefore include procedures for environment-specific baselining, re-baselining after operational changes, and careful separation between normal rare events and genuine anomalies.

At the Evaluation stage, robustness and accuracy testing are particularly important. False negatives may allow intrusions to remain undetected, while false positives may create alert fatigue or unnecessary operational concern. Evaluation should therefore test the system under normal operations, rare but legitimate industrial events, topology changes, stealthy attacks, adversarial evasion attempts, and real-time processing constraints. The system should also be tested to confirm that monitoring remains passive and does not interfere with OT processes. This corresponds directly to the evaluation block in Figure 4.1 and relies especially on prEN 18229-2 for accuracy and robustness, prEN 18282 for cybersecurity, and prEN 18228 for risk management.

At the Operation stage, the human oversight procedure should define how security teams interpret, validate, and escalate OSOTS alerts. The system should support incident response, but it should not automatically trigger operational or safety actions. Instructions for use should explain confidence levels, known limitations, expected input conditions, alert categories, validation steps, and escalation paths. Logs should record detected anomalies, model versions, network context, operator review, escalation decisions, and corrective actions. This operational design connects directly to the transparency, human oversight, logging, and post-market monitoring elements shown in Figure 4.1.

At the Retirement or change-management stage, OSOTS must be reassessed whenever it is integrated more deeply into OT processes. If future versions are connected to automated actuation, predictive maintenance decisions that materially affect safety, or safety-critical control workflows, the system would likely move into high-risk classification. The change-assessment record should therefore explicitly examine whether the system remains passive, advisory, and human-supervised.

The expected documentation outputs for OSOTS are the classification memo, intended-purpose statement, OT/ICS dataset specification, environment-specific baseline records, risk register, cybersecurity

test report, robustness and real-time performance report, instructions for use, human-oversight procedure, event logs, re-baselining procedure, and post-market monitoring plan. OSOTS can be argued to remain limited risk in its current passive configuration, but only if the system's non-actuating role and human-led incident-response model are clearly documented and maintained.

6. Overarching Assurance Pipeline for the MISSION 2 Portfolio

Across all five systems, Figure 4.1 should be understood as the common assurance pipeline for the MISSION 2 portfolio. The framework ensures that compliance is not produced only at the end of development, but is generated continuously through lifecycle activities. This is why the Provider Quality Management System under prEN 18286 is placed at the bottom of the framework: it is the governance layer that connects all AI Act obligations, lifecycle stages, standards, and documentation outputs.

At the Design stage, every MISSION 2 system should produce a classification memo and intended-purpose statement. These documents are the foundation of the entire compliance argument. They define what the system is intended to do, who will use it, what level of autonomy it has, what decisions it supports, and what it explicitly does not do. For MISSION 2, this stage is especially important because several systems could change risk classification if their purpose or autonomy level expands.

At the Development stage, each project should generate dataset specifications, data quality records, bias profiles where relevant, and training records. This stage operationalises Article 10 and ensures that data governance is not only a legal statement but a concrete engineering practice. For EPMwDC, this means controlled visual and infrared test data. For HIPSTer and AICP-FIMI, it means responsible use of public OSINT, SocMINT, and social media data. For CTI-balanced, it means secure and privacy-preserving cyber threat intelligence sharing. For OSOTS, it means representative OT/ICS traffic baselines.

At the Evaluation stage, the projects should produce risk registers, verification and validation reports, robustness tests, cybersecurity tests, and performance metrics. This stage translates Articles 9 and 15 into evidence. It should combine empirical testing with control-based verification. Empirical testing shows whether the system performs reliably in realistic conditions. Control-based verification shows whether the necessary governance, oversight, documentation, and escalation procedures are in place.

At the Operation stage, each system should have instructions for use, human oversight procedures, event logs, and post-market monitoring plans. This is where the expert-supportive character of the systems is preserved in practice. Human oversight should not be a formal statement only; it must be visible in the system interface, operational procedures, training materials, escalation rules, and logs. The operation stage is also where misuse, drift, over-reliance, false positives, and changes in deployment context must be detected.

At the Retirement and change-management stage, each project should maintain change-assessment records, version histories, decommissioning plans, and lessons learned. This stage is essential because AI Act classification is not static. A system that is limited risk today may become high-risk if it is deployed in a new context, connected to automated decisions, integrated into safety-critical processes, or used to affect individuals or democratic participation more directly.

The following table summarises how Figure 4.1 is adapted across the MISSION 2 portfolio:

Framework element from Figure 4.1	Portfolio-level application in MISSION 2
Classification and intended purpose	Each system must maintain a living classification memo defining its expert-supportive scope.
Data governance	Dataset specifications must reflect the specific data environment: visual/IR data, public OSINT/SocMINT data, social media data, CTI data, or OT/ICS traffic.
Risk management and robustness	Each system must maintain a risk register and project-specific robustness tests linked to foreseeable misuse and deployment context.
Transparency and human oversight	Outputs must remain explainable, reviewable, and subject to human decision-making.

Logging and post-market monitoring	Event logs must support traceability, incident review, model monitoring, and change assessment.
Provider QMS	prEN 18286 should coordinate classification reviews, documentation control, verification evidence, operational monitoring, and change management.

By applying Figure 4.1 in this way, Chapter 5 becomes the operational bridge between the general compliance model and the concrete MISSION 2 projects. The framework helps each system maintain its current limited- or minimal-risk profile while also preparing the portfolio for future regulatory scrutiny, sandbox participation, real-world testing, and possible evolution toward higher autonomy. Its main value is that it makes compliance visible as a lifecycle process: designed at the beginning, tested before deployment, monitored during operation, and reassessed whenever the system changes.

Conclusions and Recommendations

The EU Artificial Intelligence Act establishes a risk-based regulatory framework that demands not only legal compliance but also a disciplined engineering approach capable of producing demonstrable, auditable evidence throughout the AI lifecycle. This report has demonstrated that the five MISSION 2 AI systems – EPMwDC, HIPSTer, AICP-FIMI, CTI-balanced, and OSOTS – are intentionally designed as expert/supportive tools with limited autonomy and strong human oversight. In their current configurations, all five systems fall comfortably within the limited- or minimal-risk categories and do not trigger the full set of high-risk obligations under Annex III. This favourable classification is not accidental; it results from deliberate architectural choices, early scoping, and the systematic application of the by-design principles articulated in the Master Guide and operationalised through the Integrated EU AI Act Compliance Framework presented in Figure 4.1.

The framework provides a coherent, lifecycle-oriented roadmap that aligns each phase — Design, Development, Evaluation, Operation, and Retirement — with the relevant AI Act articles, candidate harmonised European standards, and the concrete documentation outputs required by Annex IV. By embedding the full stack of by-design principles (risk management by design, data governance and privacy by design, cybersecurity and adversarial robustness by design, human oversight and transparency by design, accountability and traceability by design, safety, resilience and continuity by design, fairness and non-discrimination by design, ethics and democratic-impact by design, and sustainability by design), the MISSION 2 projects transform regulatory requirements into natural engineering outcomes rather than administrative afterthoughts. The Provider Quality Management System (prEN 18286) serves as the unifying governance layer that ensures evidence generation is continuous, traceable, and proportionate.

Recent scholarship, particularly Buscemi et al. (2026), further reinforces this approach by offering a practical verification mapping that decomposes high-level obligations into concrete controls and testing activities across data, models, processes, and final product. Combined with the forthcoming harmonised European standards and the targeted simplifications introduced by the Digital Omnibus on AI, the MISSION 2 portfolio is exceptionally well positioned to achieve and maintain compliance while preserving its innovative character and operational effectiveness.

Key Conclusions:

1. Risk Classification Stability: All five systems remain limited or minimal risk when operated as expert/supportive tools with meaningful human oversight. Any future increase in autonomy, integration into automated decision-making, or expansion into Annex III use cases would require immediate re-evaluation and application of the full high-risk obligations; however, the by-design principles already embedded provide a robust foundation for such evolution.
2. Framework Effectiveness: The Integrated Compliance Framework (Figure 4.1) successfully translates the AI Act's horizontal requirements into a practical engineering discipline. It ensures that Annex IV technical documentation is produced organically rather than retroactively, while the by-design principles maintain both trustworthiness and regulatory alignment.
3. Digital Omnibus Synergies: The simplifications and proportionality measures introduced by the Digital Omnibus directly benefit the MISSION 2 projects. Early adoption of the by-design principles maximises these flexibilities, reduces administrative burden, and facilitates participation in regulatory sandboxes and real-world testing environments.
4. Strategic Advantage: By treating compliance as an integral part of system design from TRL 3 onward, the MISSION 2 portfolio achieves a rare combination of innovation speed, operational resilience, and regulatory readiness — positioning it as a model for secure, trustworthy, and inclusive AI applications in security-sensitive and democratic contexts.

Recommendations

General Recommendations for the MISSION 2 Programme

- Implement the Provider QMS (prEN 18286) as the central governance instrument across all projects. This single QMS should serve as the operating system for the Integrated Compliance Framework, ensuring consistent evidence generation, version control, and post-market monitoring.
- Maintain a living classification memo and risk register for each system. Update these documents whenever intended purpose, autonomy level, deployment context, or data sources change, and link them explicitly to the by-design principles.
- Adopt the by-design principles checklist from the Master Guide as a mandatory gate at every major milestone (scope approval, data readiness, model readiness, release, and post-market review).
- Engage proactively with regulatory sandboxes and real-world testing facilities under the Digital Omnibus to validate the framework in operational environments and generate additional evidence for future conformity assessments.
- Monitor the publication of harmonised European standards in the OJEU and update the technical documentation packages accordingly to benefit from the legal presumption of conformity.

Project-Specific Recommendations

- EPMwDC: Formalise the passive, preventive nature of the system in all documentation. Strengthen privacy-by-design controls for camera/IR data and maintain clear human-review gates for every alert to preserve limited-risk status.
- HIPSTer: Prioritise fairness and democratic-impact by design to avoid systematic over-flagging of protected speech. Implement analyst-facing explainability tools and robust audit trails to support post-market monitoring of hybrid-threat detection accuracy.
- AICP-FIMI: Embed ethics and democratic-impact by design at the architectural level. Ensure all detection outputs remain advisory and include explicit uncertainty signalling to election monitors, thereby avoiding any shift toward high-risk automated content moderation.
- CTI-balanced: Reinforce data governance and privacy by design through strong anonymisation and consent mechanisms. Document the balanced-incentive model thoroughly to demonstrate that the platform supports collective cybersecurity without individual profiling.
- OSOTS: Explicitly document the passive monitoring character and human-in-the-loop incident-response workflow. Prepare a detailed safety and resilience justification to support the Article 6(3) derogation should the system be deployed in critical infrastructure environments.

By following these conclusions and recommendations, the MISSION 2 consortium will not only achieve full regulatory compliance under both the EU AI Act and the Digital Omnibus but will also deliver trustworthy, human-centric AI solutions that meaningfully contribute to a safe and inclusive e-society. The integrated framework, by-design principles, and proactive governance approach described throughout this report provide a replicable model for other European R&D initiatives seeking to balance innovation with responsibility in high-stakes domains.

Literature

1. Nannini, L., Smith, A. L., Maggini, M. J., Panai, E., Feliciano, S., Tiulkanov, A., Maran, E., Gealy, J., & Bisconti, P. (2026). *AI agents under EU law: A compliance architecture for AI providers*. arXiv. https://arxiv.org/abs/2604.04604
2. Buscemi, A., Deckenbrunnen, T., Kabir, F., Mishchenko, K., & Mowla, N. (2025). *Assessing high-risk AI systems under the EU AI Act: From legal requirements to technical verification*. arXiv. https://arxiv.org/abs/2512.13907
3. AI Act Standards. (n.d.). *Interactive mapping of AI Act standards*. https://ai-act-standards.com/
4. CEN-CENELEC Joint Technical Committee 21 (Artificial Intelligence). (n.d.). *Standards development for artificial intelligence*. https://standards.cencenelec.eu/ords/f?p=205:22:::::FSP_ORG_ID,FSP_LANG_ID:2916257,25&cs=114251C6C0B684FBBC069923513BF6348
5. Artificial Intelligence Act. (n.d.). *Official website of the EU Artificial Intelligence Act*. https://artificialintelligenceact.eu/
6. Council of the European Union. (2026, March 13). *Council agrees position to streamline rules on artificial intelligence* [Press release]. https://www.consilium.europa.eu/en/press/press-releases/2026/03/13/council-agrees-position-to-streamline-rules-on-artificial-intelligence/

Finansuoja Europos Sąjunga
NextGenerationEU

Mykolo Romerio universitetas

NAUJOS KARTOS LIETUVA

Editor
Valentas Gaižauskas (MRU)

Contributors
Valentas Gaižauskas (MRU)
Evaldas Bruze (MRU)
Giedrė Sabaliauskaitė (MRU)
R. Andrew Paskauskas (MRU)
Tomas Lavišius (MRU)

Reviewers
Valentas Gaižauskas (MRU)
Evaldas Bružė (MRU)
Giedrė Sabaliauskaitė (MRU)
R. Andrew Paskauskas (MRU)
Tomas Lavišius (MRU)

Disclaimer