Finansuoja Europos Sąjunga
NextGenerationEU

Mykolo Romerio universitetas

NAUJOS KARTOS LIETUVA

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

Project name: ***AICP-FIMI***

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

Action result: ***Architecture: Incorporating Regulatory and Security Requirements in the AICP-FIMI Platform***

Architecture: Incorporating Regulatory and Security Requirements in the AICP-FIMI Platform

Table of Contents

Introduction

The project “AI driven cloud platform to counter FIMI during elections and early warning service for identification of social media bot and troll farms (AICP-FIMI)” focuses on the development of an AI-driven cloud platform designed to combat Foreign Influence and Manipulation of Information (FIMI) during elections in Lithuania and other countries. The platform will provide a service to identify social media bots and troll farms, enhancing the nation's resilience against cyber threats. By providing accurate and timely information, the project aims to strengthen democratic processes and promote resistance to cyber security issues.

The automation of detecting social media bots and troll farms faces challenges such as rapidly evolving tactics and the dynamic nature of social media platforms. Developing an effective automated system requires constant updates and adaptations to identify and mitigate emerging threats accurately.

In implementing this project, it is crucial to examine recent policy practices related to the implementation of the GDPR and AI in the field, as well as the case law of the Court of Justice of the EU related to data protection and automated data processing.

This document provides a comprehensive analysis of the Architecture: Incorporating Regulatory and Security Requirements in the AICP-FIMI Platform.

1. Core Architectural Principles: Privacy and Security by Design

The architecture of the AI-driven Cloud Platform for Foreign Influence and Manipulation of Information (AICP-FIMI) must be built upon a foundation of strict regulatory compliance, particularly the GDPR, the EU AI Act, the NIS2 Directive, and the Cyber Resilience Act (CRA).

- **Privacy by Design and Default (GDPR Art. 25):** Privacy considerations must be embedded into the platform’s architecture from the initial conceptual phase, ensuring that data minimization, anonymization, and pseudonymization techniques are applied throughout the entire data processing lifecycle.
- **Security by Design and Default (CRA & NIS2):** The framework must integrate security measures (such as vulnerability handling) from the earliest design phase rather than adding

them as an afterthought. The architecture must implement robust **Role-Based Access Controls (RBAC)** to restrict data access exclusively to authorized personnel.

2. Key Architectural Layers and Components

The platform is designed as a modular and scalable cloud-based architecture to allow flexible deployment across different regions and handle high data volumes during elections.

- **Cloud Infrastructure:** The system must utilize **elastic compute resources** to dynamically scale processing power, backed by **secure data storage** containing encryption and audit logs. It must also incorporate redundancy and disaster recovery mechanisms to maintain continuous operation.
- **Data Ingestion Layer:** This layer collects data in real-time from social media platforms and news websites using APIs and web scraping techniques. Architecturally, it must apply data minimization filters directly at the point of collection to exclude biometric data (like profile photos) and indiscriminately scraped personal identifiers.
- **Processing Layer:** The core utilizes machine learning and Natural Language Processing (NLP) models to identify disinformation patterns (e.g., topic modeling, sentiment analysis, entity recognition).
- **Analysis and Detection Layer:** This layer employs **graph-based network analysis** to detect coordinated behavior. It must incorporate algorithms for **Community Detection** (to find troll farms), **Centrality Measures** (to find influential nodes), and **Temporal Analysis** (to track campaign evolution).
- **Alert and Reporting Layer:** Generates customizable alerts and incorporates a comprehensive, user-friendly dashboard for visualizing data patterns and accessing detailed reports.

3. Agentic Architecture & Protocol Compliance

To prevent "Black Box" liability and ensure interoperability, the system must standardize agent communication and tool use.

- **Tool Interfaces (MCP):** Autonomous agents must connect to external tools (like social media APIs or databases) via the **Model Context Protocol (MCP)** to standardize security borders. The architecture must enforce "User Consent" prompts before an agent executes any "Write" action, such as reporting a bot.
- **Inter-Agent Communication (A2A):** The architecture should use the **Agent2Agent (A2A) protocol** for coordination between different sub-agents (e.g., "Research Agents" and "Analysis Agents"). This requires creating "Agent Cards" (JSON metadata) defining capabilities and tracking task lifecycles.
- **Identity & Access (Non-Human Identities):** Agents must be treated as Non-Human Identities (NHIs) with granular permissions. For example, Research Agents should have "Read-Only" access, while "Human-in-the-Loop" (HITL) tokens are required for high-risk actions.
- **Agentic RAG (Retrieval-Augmented Generation):** The architecture must move beyond static retrieval to "Iterative Retrieval," allowing agents to self-correct queries rather than hallucinate. It must also enable **"Citation Matching"**, ensuring the model links every generated claim to a specific document ID.

4. Algorithmic Transparency & Explainability (XAI)

To comply with the GDPR's "Right to Explanation" (Art. 15 and 22) and CJEU case law (Schufa, Dun & Bradstreet), the architecture cannot function as a "black box".

- **Chain of Thought (CoT) Logging:** The orchestration layer must log the reasoning steps (Query -> Plan -> Tool Call -> Observation -> Conclusion) and store these logs in an **immutable audit trail**.
- **Logic Disclosure:** The User Interface (UI) must explicitly display why an account was flagged (e.g., "Flagged because posting frequency > 500/hr AND network centrality > 0.9") rather than assigning a generic "AI Score".
- **Hallucination Checks:** The architecture must include a **"Verifier Agent"** a secondary model designed to perform a Grounding Check, ensuring that the generated reports accurately match the retrieved source data.

5. Operational Governance and Cybersecurity Architecture

The architecture must be reinforced by the **TRAPS (Trusted, Responsible, Auditable, Private, Secure)** framework to manage the agent lifecycle and cybersecurity.

- **Data Security and Encryption:** All data must be encrypted at rest (e.g., using **AES-256**) and in transit (e.g., using **TLS 1.2+**). Output guardrails must execute **PII Redaction** to strip names and IDs before reports leave the secure enclave.
- **Incident Detection and Response:** The architecture must feature **Intrusion Detection Systems (IDS)** using anomaly detection models to monitor network traffic. Automated response protocols must be in place to contain threats.
- **Circuit Breakers:** The system must incorporate automated stops ("Circuit Breakers") that halt an agent if it exceeds API rate limits or accesses unauthorized endpoints, mitigating the risk of runaway agents.
- **Human-in-the-Loop (HITL) Workflow:** The deployment architecture must operate as a **"Cyborg" workflow (Human-AI hybrid)**, heavily relying on human oversight for high-risk decisions (like automated takedowns), ensuring humans handle final judgment while agents manage data scale.

Annex: Architecture

- **Stages:**
  - **Data Ingestion Layer:** Collecting data in real-time using APIs and web scraping, where data minimization filters must be applied.
  - **Processing Layer:** Utilizing Machine Learning and Natural Language Processing (NLP) to parse ingested data.
  - **Analysis and Detection Layer:** Employing graph-based network mapping (community detection, centrality measures) to identify coordinated behavior and troll farms.
  - **Alert and Reporting Layer:** Generating alerts and actionable dashboards for authorities.
- **Compliance Checklist:**
  - **Privacy and Security by Design:** Encryption implemented for data at rest (AES-256) and in transit (TLS 1.2+), alongside Role-Based Access Controls (RBAC).

- **Standardized Agent Interfaces:** Model Context Protocol (MCP) servers used for all data connectors, and Agent2Agent (A2A) protocol implemented for inter-agent communication and task tracking.
- **Algorithmic Transparency (XAI):** "Chain of Thought (CoT) Logging" enabled in the orchestration layer, storing the reasoning steps from Query to Conclusion in an immutable audit trail.
- **Hallucination Checks:** Integration of a "Verifier Agent" to conduct grounding checks, ensuring generated reports strictly match retrieved source data.
- **TRAPS Framework implemented:** Operational controls mapped to Trusted, Responsible, Auditable, Private, and Secure principles, including "Circuit Breakers" to halt runaway agents.

- **Challenges:**
  - **Explainability of "Black Box" Models:** Modern deep learning models act as "black boxes," making it technically difficult to articulate their logic. This directly conflicts with the GDPR "Right to Explanation" and the CJEU *Dun & Bradstreet* ruling.
  - **Cross-Border Regulatory Complexity:** Creating an architecture that simultaneously harmonizes with the GDPR, NIS2 Directive, DSA, and potential NATO operational frameworks is highly complex.
- **Suggestions to Overcome Challenges:**
  - Design the User Interface (UI) to display explicit logic (e.g., "Flagged because posting frequency > 500/hr AND network centrality > 0.9") rather than generic "AI Scores".
  - Move from static retrieval to **Agentic RAG with "Citation Matching"**, ensuring the system links every generated claim to a specific document ID for immediate provenance and explainability.
  - Develop a **unified data protection protocol** to establish consistent cybersecurity and compliance practices that meet both EU and national requirements across all operational contexts.

Conclusions

The architectural framework of the AICP-FIMI platform must balance advanced technological capabilities, such as Natural Language Processing (NLP) and graph-based network analysis, with strict legal mandates for transparency and security. Operating as a "black box" is legally unviable; the architecture must inherently support Explainable AI (XAI) and "Chain of Thought" logging to ensure all automated decisions can be audited and understood by human operators.

Additional Insights

- **The Transparency Imperative:** Based on CJEU jurisprudence (specifically the *Schufa* and *Dun & Bradstreet* cases), if the platform's architecture flags an account and it leads to significant effects (like automated censorship or account suspension), it constitutes "automated individual decision-making". Providers cannot use trade secrets as a blanket excuse to hide their algorithms; the system's User Interface must provide "meaningful information about the logic involved" (e.g., detailing the specific parameters that led to a bot classification).
- **Managing Hybrid System Complexity:** The platform's architecture will likely be a "hybrid system" combining AI components with non-AI digital elements. Under the Cyber Resilience Act (CRA), this requires continuous, holistic risk management across all components, because a vulnerability introduced during AI training could manifest as a critical security flaw in a downstream application.
- **Cross-Border Data Sharing Challenges:** The architecture must incorporate **Unified Data Protection Protocols** to navigate the complexities of international operations. GDPR imposes strict conditions on transferring personal data outside the EU, which complicates cross-border data sharing, especially when trying to align EU legal requirements with NATO's strategic operational frameworks and intelligence-sharing needs.

References

1. Artificial Intelligence Act: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401689
2. General Data Protection Regulation: https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng
3. CJEU Case C-446/21 - Meta Platforms Ireland, 2024. https://curia.europa.eu/juris/document/document.jsf?text=&docid=290674&pageIndex=0&doclang=EN&mode=lst&dir=&occ=first&part=1&cid=8907758
4. CJEU Case C-203/22 - Dun & Bradstreet Austria GmbH, 2025. https://curia.europa.eu/juris/liste.jsf?num=C-203/22
5. CJEU Case C-634/21 - SCHUFA Holding AG, 2023. https://curia.europa.eu/juris/liste.jsf?num=C-634/21
6. CJEU Case C-511/18 - La Quadrature du Net and Others, 2020. https://curia.europa.eu/juris/liste.jsf?language=en&num=C-511/18
7. Digital Services Act. https://eur-lex.europa.eu/eli/reg/2022/2065/oj/eng
8. The Cyber Resilience Act: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202402847
9. NIS2 Directive: https://eur-lex.europa.eu/eli/dir/2022/2555/2022-12-27/eng
10. CJEU Case C-413/23 - EDPS v SRB, 2025. https://curia.europa.eu/juris/document/document.jsf?text=&docid=305584&pageIndex=0&doclang=EN&mode=req&dir=&occ=first&part=1&cid=8121207
11. CJEU Case C-817/19 - Ligue des droits humains. https://curia.europa.eu/juris/liste.jsf?num=C-817/19
12. Cyber Solidarity Act: https://eur-lex.europa.eu/eli/reg/2025/38/oj/eng
13. Case C-250/25 - Like Company v. Google Ireland Ltd. https://curia.europa.eu/juris/liste.jsf?num=C-250/25
14. Fundamentals of Secure AI Systems with Personal Data. EDPB Guidance. April 2025. https://www.edpb.europa.eu/system/files/2025-06/spe-training-on-ai-and-data-protection-technical_en.pdf
15. Guidance for Risk Management of Artificial Intelligence systems. European Data Protection Supervisor. November 2025. https://www.edps.europa.eu/data-protection/our-

work/publications/guidelines/2025-11-11-guidance-risk-management-artificial-intelligence-systems_en

16. Guidance on Generative AI, strengthening data protection in a rapidly changing digital era. European Data Protection Supervisor. October 2025. https://www.edps.europa.eu/data-protection/our-work/publications/guidelines/2025-10-28-guidance-generative-ai-strengthening-data-protection-rapidly-changing-digital-era_en
17. First EDPS Orientations for EUIs using Generative AI. European Data Protection Supervisor. June 2024. https://www.edps.europa.eu/data-protection/our-work/publications/guidelines/2024-06-03-first-edps-orientations-euis-using-generative-ai_en
18. Guidelines on the scope of obligations for providers of General-Purpose AI (GPAI) models. July 2025. https://digital-strategy.ec.europa.eu/en/library/guidelines-scope-obligations-providers-general-purpose-ai-models-under-ai-act
19. Guidelines on the AI system definition. February 2025. https://digital-strategy.ec.europa.eu/en/library/commission-publishes-guidelines-ai-system-definition-facilitate-first-ai-acts-rules-application
20. General-Purpose AI Code of Practice. July 2025. https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai
21. First Draft Code of Practice on Transparency of AI-Generated Content (December 2025). https://ec.europa.eu/newsroom/dae/redirection/document/123074

Editor
<<Main Editor>>

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

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

Disclaimer