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: ***System Analysis: Incorporating GDPR and Regulatory Requirements in the AICP-FIMI Project Lifecycle***

System Analysis: Incorporating GDPR and Regulatory Requirements in the AICP-FIMI Project Lifecycle

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 System Analysis. The successful implementation of the AI-driven cloud platform to counter Foreign Influence and Manipulation of Information (AICP-FIMI) heavily depends on ensuring compliance with the core principles of the General Data Protection Regulation (GDPR) and the EU AI Act throughout the project lifecycle. Because the platform will monitor social media to detect bot farms and manipulation during elections, it inherently processes large volumes of personal data and qualifies as a High-Risk AI System. System analysis must establish robust architectural, technical, and operational safeguards to oversee the secure collection and processing of personal data.

1. Core GDPR Principles to Integrate During System Analysis

During the system analysis phase, the architecture must be designed to strictly adhere to the following GDPR principles:

- **Privacy by Design and by Default (Article 25):** Security and privacy measures must be integrated from the earliest conceptual phase, not added as an afterthought. System analysis must define mechanisms for **anonymization and pseudonymization** to protect individual identities. However, strict access controls are necessary because pseudonymized data may still be considered personal data if re-identification is possible.
- **Data Minimization and Storage Limitation (Article 5(1)(c)):** The system must collect only data that is "strictly necessary" for bot detection. System analysis must include data filtering at the point of collection to prevent indiscriminate data aggregation. Technical controls such as **"Time-To-Live" (TTL) on all raw scraped data** and **"Semantic Caching"** (storing abstract patterns instead of raw user data) must be implemented. General and indiscriminate retention of data is legally prohibited unless there is a present, serious threat to national security.
- **Security of Processing (Article 32):** The system must employ appropriate technical and organizational measures. This includes **encrypting all data at rest and in transit** (e.g., using AES-256 for storage and TLS 1.2+ for transmission) and implementing **Role-Based Access Controls (RBAC)** to restrict data access to authorized personnel.
- **Transparency and Explainable AI (XAI) (Articles 15 & 22):** If the platform's identification of a "bot" leads to significant automated effects (e.g., account suspension), it constitutes automated decision-making. Providers must supply "meaningful information about the logic involved". The system must be able to generate reports explaining exactly why an account was flagged (e.g., “posting frequency > 500/hr AND network centrality > 0.9”) and avoid generic "AI Score" labels.

2. Specific Requirements and Regulatory "Red Lines"

The system analysis must establish clear technical guardrails to prevent the system from executing prohibited practices (unacceptable risks) under the AI Act and GDPR:

- **Prohibition of Biometric Scraping:** The platform must strictly avoid untargeted scraping of facial images from the internet or social media to build recognition databases. Web scrapers must be programmed to explicitly exclude .jpg/.png collections from user profiles.

- **No Political Profiling via Biometrics:** The system cannot use biometric data to categorize political opinions. Classification models must rely solely on behavioral metadata (e.g., posting frequency, network patterns) and text semantics, never on physical attributes.
- **Secure Data Acquisition:** Indiscriminate web scraping carries high legal risks under the GDPR. The system should be designed to utilize the lawful data access mechanism provided by the **Digital Services Act (DSA)** framework, allowing vetted researchers to access data from Very Large Online Platforms (VLOPs) to analyze systemic risks securely.

3. Key Implementation Stages and System Aspects

To ensure end-to-end compliance, the system analysis must break down the project lifecycle into four critical stages, incorporating specific requirements at each step:

Stage 1: Design and Impact Assessment Phase

- **Assessments:** Before data processing begins, the project must conduct a **Data Protection Impact Assessment (DPIA)** and a **Fundamental Rights Impact Assessment (FRIA)** to evaluate risks to democracy, non-discrimination, and privacy.
- **Logic Definition:** The analysis must clearly define the behavioral markers for "bot" identification rather than political content to prevent prohibited profiling.
- **Cybersecurity Framework:** Implement a "Security-by-Design" framework aligned with the NIS2 Directive and Cyber Resilience Act, defining incident detection and response protocols.

Stage 2: Data Acquisition and Processing Phase

- **Secure Interfacing:** Define Tool Interfaces using the Model Context Protocol (MCP) to standardize security borders for data connectors (e.g., social media APIs).
- **Data Filtration:** Apply data minimization filters directly at the point of ingestion to exclude biometric data and unnecessary personal identifiers.
- **TRAPS Framework implementation:** Operational governance must follow the **TRAPS (Trusted, Responsible, Auditable, Private, Secure)** framework. This includes establishing "Circuit Breakers" to automatically stop agents exceeding rate limits, and implementing

Output Guardrails to redact Personally Identifiable Information (PII) before reports leave secure enclaves.

Stage 3: Model Training and Testing Phase

- **Adversarial Testing (Red-Teaming):** The system must undergo "red-teaming" to test models against "data poisoning," "prompt injection" attacks, and biases.
- **Bias Mitigation:** Special categories of data (like political orientation) can be processed *only* within the strict confines of bias monitoring and correction to ensure the model does not unfairly target specific demographic groups.
- **Transparency Markers:** The models must be trained to recognize technical transparency markers (e.g., C2PA metadata) to detect AI-generated content.

Stage 4: Operational Deployment and Oversight Phase

- **Human-in-the-Loop (HITL) Oversight:** The system architecture must incorporate a mandatory human review process. A human must verify the AI's classification before taking high-risk actions (such as alerting the public or reporting for takedown), aligning with a **"Cyborg" workflow (Human-AI hybrid)**.
- **Auditability:** The system must implement "Chain of Thought (CoT) Logging," tracking the entire reasoning process from Query to Conclusion, stored in an immutable audit trail.
- **Hallucination Prevention:** The deployment phase must include a "Verifier Agent" to cross-check if the generated reports accurately match the retrieved source data, preventing the fabrication of evidence.

Annex: System Analysis

- **Stages:**
    - **Design and Impact Assessment Phase:** This initial conceptual phase involves establishing the legal basis, architectural safeguards, and conducting mandatory risk assessments.
    - **Data Acquisition and Processing Phase:** The stage focused on securely and lawfully collecting operational data and applying minimization filters.

- **Model Training and Testing Phase:** Focuses on training the models while ensuring robust defenses against bias and adversarial attacks (red-teaming).

- **Compliance Checklist:**
  - **Data Protection Impact Assessment (DPIA)** and **Fundamental Rights Impact Assessment (FRIA)** conducted to evaluate risks to privacy, non-discrimination, and democracy.
  - **Prohibited practices actively avoided:** System programmed to exclude the untargeted scraping of facial images and the use of biometric data for inferring political opinions.
  - **Data Minimization verified:** "Time-To-Live" (TTL) on all raw scraped data and data filtering at the point of ingestion to exclude unnecessary personal identifiers.
  - **Logic Definition:** Behavioral markers (e.g., posting frequency) clearly defined for "bot" identification rather than relying on political content.
- **Challenges:**
  - **Legal Ambiguity of "Subliminal Techniques":** Distinguishing between a sophisticated FIMI bot and a passionate human activist using persuasive language creates a risk of false positives that could infringe upon freedom of speech.
  - **Data Access vs. Scraping:** Relying on web scraping carries severe legal risks under the GDPR (lack of legal basis) and the AI Act (prohibited biometric scraping).
- **Suggestions to Overcome Challenges:**
  - Instead of indiscriminate scraping, utilize the lawful data access mechanism provided by the **Digital Services Act (DSA) framework**, which allows vetted researchers to analyze systemic risks from Very Large Online Platforms (VLOPs) securely.
  - Ensure models rely solely on **technical behavioral metadata** (e.g., posting frequency, network patterns, Actor/Behavior/Content frameworks) and text semantics, strictly avoiding biometric or content-based political profiling

Conclusions

The system analysis phase of the AICP-FIMI project must navigate a complex regulatory "minefield" of prohibited practices and high-risk classifications established by the EU AI Act and the GDPR. The successful conceptualization of the platform requires integrating Privacy by Design from the very beginning. It is mandatory to conduct Data Protection Impact Assessments (DPIAs) and Fundamental Rights Impact Assessments (FRIAs) before any data is processed to evaluate the potential risks to democracy, non-discrimination, and individual freedoms.

Additional Insights

- **Strategic Alignment and Urgency:** The system analysis is supported by a strong political mandate. Initiatives like the "European Democracy Shield" explicitly call for tools to trace "coordinated inauthentic behaviour" and "bot-driven amplification," validating the urgent operational need for this platform to protect electoral integrity.
- **Lawful Data Access over Web Scraping:** The AI Act strictly prohibits the untargeted scraping of facial images to build recognition databases and bans the use of biometric data to infer political opinions. A key insight for system analysis is to avoid legally risky indiscriminate web scraping; instead, the project should leverage the lawful data access mechanisms provided by the **Digital Services Act (DSA)** to access data from Very Large Online Platforms (VLOPs) for analyzing systemic risks.
- **Mitigating Negligence and Liability:** The CJEU ruling in *Deutsche Wohnen* established that administrative fines under the GDPR can only be imposed if an infringement is due to negligent or willful conduct. Thorough system analysis, well-documented risk mitigation, and strict adherence to data minimization (e.g., implementing Time-to-Live limits on data) are crucial to prove an absence of negligence in the event of a cybersecurity breach.

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