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: ***Deployment: Regulatory and Operational Requirements for the AICP-FIMI Platform***

Deployment: Regulatory and Operational Requirements for 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 Deployment phase of the AI-driven Cloud Platform for Foreign Influence and Manipulation of Information (AICP-FIMI), which marks the critical transition from development to live, operational environments. During this stage, the system begins actively processing real-world data to identify social media bots and troll farms. To ensure strict compliance with the General Data Protection Regulation (GDPR), the EU AI Act, the Digital Services Act (DSA), and existing case law, deployment must be managed as an ongoing process of rigorous oversight, human-machine collaboration, and continuous monitoring.

1. The Deployment and "Automatability" Roadmap

Deployment must follow a phased approach to align the project timeline with the maturity of compliance-automating technologies and to mitigate high-risk legal exposures.

- **Phase 1: Automatability Trigger (Design Control):** The platform must not activate any fully automated takedown or alerting features until "Compliance-Automating AI" (agents capable of checking other agents against EU rules) is fully integrated. This reduces the risk of premature regulatory violations and minimizes high human oversight costs.

- **Phase 2: Hybrid Teaming (Deployment):** The system must be deployed as a **"Cyborg" workflow (Human-AI hybrid)** rather than a fully autonomous entity. In this operational model, AI agents handle the massive scale of data processing, while human operators handle judgment and final decisions. Studies suggest this hybrid deployment outperforms fully autonomous agents in quality by 68.7%.
- **Phase 3: Continuous Monitoring:** Once live, the deployment must actively utilize **Guardrails** (e.g., NeMo or simple deterministic rules) to filter system inputs and outputs. This protects the operational platform against adversarial attacks, such as "Prompt Injections" or "Jailbreaks", which attempt to weaponize the FIMI platform.

2. Specific Requirements and Key Aspects During Deployment

To maintain regulatory compliance during live operations, several specific mechanisms must be active and strictly enforced:

- **Human-in-the-Loop (HITL) Oversight:** A non-negotiable requirement for deployment is mandatory human oversight. A human must verify the AI's classification of a "bot" before the system takes any significant or high-risk action, such as reporting an account to a social media platform for a takedown or issuing public alerts. This is a direct requirement stemming from CJEU case law (specifically *Ligue des droits humains*), which mandates that automated matches be subject to individual human review.
- **Explainable AI (XAI) and Logic Disclosure:** The platform must not operate as a "black box" in deployment. It must actively use Explainable AI (XAI) mechanisms to provide transparency into how decisions are made. According to CJEU rulings (like *Dun & Bradstreet*), the system must generate detailed reports explaining the specific "logic involved" when an account is flagged (e.g., clarifying that an account was flagged because it "posted 500 times in 1 hour").
- **Public Feedback and Mechanisms for Redress:** At the time of launch, the platform must have active systems in place to allow the public to provide feedback and seek redress. If an individual is incorrectly flagged by the system as a bot or part of a troll farm, they must have an accessible, participatory mechanism to contest the decision, which is vital for building public legitimacy and trust.
- **Incident Reporting and Ecosystem Integration:** The deployed platform must be operationally integrated with broader European cybersecurity ecosystems. It should connect with the **Cyber Solidarity Act** ecosystem to report detected large-scale FIMI incidents or cyber threats. Furthermore,

to comply with the NIS2 Directive, automated protocols must be active to report any significant internal cybersecurity incidents directly to relevant national authorities within specified timeframes.

- **Launch and Continuous Compliance (Months 19-36):** The actual launch of the platform must occur only with full compliance measures in place. Post-launch, the deployment phase requires continuous operational governance, including:
    - Conducting regular compliance audits and security updates.
    - Generating and publishing regular **Transparency Reports**, a key obligation under the DSA (Article 23), detailing the actions taken against disinformation to ensure public accountability.
    - Continuously monitoring and updating risk management processes and ethical guidelines to adapt to evolving FIMI tactics.

Annex: Deployment

- **Stages:**
    - **Phase 1: Automatability Trigger (Design Control):** Restricting fully automated takedown features until "Compliance-Automating AI" is fully integrated to prevent premature regulatory violations.
    - **Phase 2: Hybrid Teaming (Deployment):** Going live with a "Cyborg" workflow (Human-AI hybrid) where AI processes data scale and humans handle judgment.
    - **Phase 3: Continuous Monitoring:** Utilizing guardrails and regular audits post-launch to actively filter out adversarial attacks.
- **Compliance Checklist:**
    - **Human-in-the-Loop (HITL) Oversight:** Human verification mandated before any high-risk action (e.g., public alerts or platform takedown reporting) is taken, complying with the *Ligue des droits humains* ruling.
    - **Incident Reporting:** Active integration with the NIS2 Directive and Cyber Solidarity Act ecosystems to promptly report significant cybersecurity incidents to national authorities.
    - **Transparency Reporting:** Mechanisms in place to regularly publish DSA-mandated reports detailing the actions taken against disinformation.
    - **Public Feedback and Redress:** Systems operational to allow the public to contest decisions (e.g., if a human is incorrectly flagged as a bot farm participant).
- **Challenges:**

- **The Scale vs. Nuance Tradeoff:** Heavy reliance on human oversight (to ensure nuance and compliance) limits the speed and scalability needed to match highly automated bot networks.
- **Public Perception Risks:** The platform risks being perceived by the public as an instrument of state surveillance or censorship, which could polarize society further.
- **Adversarial Weaponization:** During deployment, the platform faces constant threats from "prompt injections," "jailbreaks," and data poisoning from state-aligned actors attempting to exploit the system.

- **Suggestions to Overcome Challenges:**
  - Utilize the **Hybrid Teaming ("Cyborg") model**, as studies indicate hybrid teams outperform fully autonomous agents by 68.7% in quality. Let the AI agents handle the vast data mapping while dedicating human analysts strictly to high-stakes validation.
  - Actively deploy **Input/Output Guardrails (e.g., NeMo)** and conduct ongoing "red-teaming" to filter out malicious prompts and protect against adversarial weaponization.
  - To mitigate negative public perception, maintain a transparent, **"whole-of-society" approach**, engaging openly with civil society, media, and educational institutions, and ensure redress mechanisms are highly accessible.

Conclusions

Deployment is not a static endpoint but an ongoing phase of continuous monitoring, adversarial testing, and mandatory human oversight. To comply with high-risk AI obligations and CJEU case law (specifically *Ligue des droits humains*), the platform must be deployed as a "Cyborg" workflow (a human-AI hybrid). Any automated match or classification generated by the AI must be subject to an individual human review before any high-risk action, such as public alerting or reporting for takedown, is executed.

Additional Insights

- **Ecosystem Integration:** The deployed platform should not operate in an isolated silo. To maximize its early-warning capabilities, it is highly recommended to integrate the platform with the broader EU security ecosystem, such as the **Rapid Response System** used during elections, the Cyber Solidarity Act's alert system, and the FIMI Information Sharing and Analysis Centre (FIMI ISAC).

- **Addressing Generative AI Risks:** As malicious actors increasingly use generative AI for FIMI, the deployment phase must actively monitor for "hallucinations" and defend against technical vulnerabilities like "prompt injections" or "jailbreaks". Deployment must include ongoing "red-teaming" (adversarial testing) and utilize technical markers (such as C2PA metadata) to detect and label AI-generated content.
- **The "Whole-of-Society" Approach:** Technological deployment alone cannot defeat disinformation. The operational phase must be coupled with a "whole-of-society approach" that engages civil society, educational institutions, and the media. This is necessary to build public awareness and media literacy, creating psychological resilience against the "Fear, Uncertainty, and Doubt" tactics heavily utilized in hybrid warfare campaigns.

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)