European Approach to Regulating AI

Adrian Groza

Cadru strategic pentru adoptarea si utilizarea de tehnologii inovative in administratia publica 2021-2027– solutii pentru eficientizarea activitatii
COD: SIPOCA 704
Beneficiar: AUTORITATEA PENTRU DIGITALIZAREA ROMANIEI
Partener: UNIVERSITATEA TEHNICA DIN CLUJ-NAPOCA

European approach to Artificial intelligence (Link)

- human centric: centres on excellence and trust; aims to boost research and industrial capacity and ensure fundamental rights (AI should be a tool for people)
- addresses the opacity, complexity, bias, a certain degree of unpredictability and partially autonomous behaviour
- expects higher demand due to higher trust, more available offers due to legal certainty, and the absence of obstacles to cross-border movement of AI systems

Objectives

1. ensure that AI systems placed on the Union market and used are safe and respect existing law on fundamental rights and Union values;
2. ensure legal certainty to facilitate investment and innovation in AI;
3. enhance governance and effective enforcement of existing law on fundamental rights and safety requirements applicable to AI systems;
4. facilitate the development of a single market for lawful, safe and trustworthy AI applications and prevent market fragmentation.

Selecting among 5 options

- O_1 EU legislative instrument setting up a voluntary labelling scheme;
- O_2 A sectoral, “ad-hoc” approach;
- O_3 Horizontal EU legislative instrument following a proportionate risk- based;
- O_{3+} Horizontal EU legislative instrument following a proportionate risk- based approach + codes of conduct for non-high-risk AI systems;
- O_4 Horizontal EU legislative instrument establishing mandatory requirements for all AI systems, irrespective of the risk they pose.

AI techniques and approaches

1. Machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning;
2. Logic-and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems;
3. Statistical approaches, Bayesian estimation, search and optimization methods.

Three clusters of AI systems

1. Prohibited AI practices (an unacceptable risk): 4 practices
2. High-Risk AI Systems (HRAIS): 8 domains
3. Minimal-risk AI systems (MRAIS)

Prohibited AI practices (an unacceptable risk)

1. AI deploying subliminal techniques beyond a person’s consciousness in order to materially distort a personâs behaviour in a manner that causes or is likely to cause that person or another person physical or psychological harm;
2. AI exploiting any of the vulnerabilities of a specific group of persons due to their age, physical or mental disability, in order to materially distort the behaviour of a person pertaining to that group in a manner that causes or is likely to cause that person or another person physical or psychological harm;
3. AI systems by public authorities or on their behalf for the evaluation or classification of the trustworthiness of natural persons over a certain period of time based on their social behaviour or known or predicted personal or personality characteristics (i.e. social scoring)
4. the use of “real-time“ remote biometric identification systems in publicly accessible spaces (exceptions, e.g., the targeted search for specific potential victims of crime, including missing children)

High-risk AI systems I

1. Biometric identification and categorisation of natural persons: “real-time”/“post”
2. Operation of critical infrastructure: traffic, supply of water, gas, heating, electricity
3. Education and vocational training: determining access or assigning natural persons to educational and vocational training institutions
4. Employment, workers management and access to self-employment:
   - recruiting, notably for advertising vacancies, screening or filtering applications, evaluating candidates in the course of interviews or tests;
   - making decisions on promotion and termination of work contracts, for task allocation and for monitoring and evaluating performance and behavior
5. Access to and enjoyment of essential private and public services and benefits:
   - public authorities to evaluate the eligibility of natural persons for public assistance benefits (e.g. reduce, revoke, or reclaim)
   - to evaluate the creditworthiness of natural persons (with some exceptions)
6. Law enforcement
   - Assessing the risk of a natural person for offending or reoffending, or the risk for potential victims of criminal offences;
   - AI as polygraphs and similar tools or to detect the emotional state;

High-risk AI systems II

- AI systems used by law enforcement authorities to detect deep fakes
- Evaluating of the reliability of evidence in the course of investigation or prosecution of criminal offences;
- Predicting the occurrence or reoccurrence of an actual or potential criminal offence based on profiling of natural persons
- Profiling during detection, investigation or prosecution of criminal offences;
- Crime analytics regarding natural persons (i.e. datamining)

7. Migration, asylum and border control management
    - AI polygraphs and similar tools to detect the emotional state;
    - Assessing a risk (e.g., security, irregular immigration, health) posed by a natural person who intends to enter into the territory of a Member State;
    - Verification of the authenticity of travel and supporting documentation and detect non-authenticity by checking their security features;
    - Assisting competent public authorities for the examination of applications for asylum, visa and residence permits and associated complaints with regard to the eligibility of the natural persons applying for a status.
8. Administration of justice and democratic processes: assisting a judicial authority in researching and interpreting facts and the law, and in applying the law

Minimal-risk AI systems

Transparency obligations

- flag the use of an AI system when interacting with humans
- disclose content generated through automated means (e.g., deep fakes)
- operationalised through harmonised technical standards

Other aspects

- Minimise the risk of algorithmic discrimination: design and the quality of data sets, obligations for testing, risk management, documentation and human oversight throughout the AI systems' lifecycle.
- Expertise for auditing is only now being accumulated – use expertise for products covered by the New Legislative Framework (NLF) legislation (e.g. machinery, medical devices, toys, lifts, equipment and protective systems; financial services legislation (e.g., credit institutions)
- Conformity assessment procedures to be followed for each high-risk AI system
- Re-assessments in case of substantial modifications to AI systems

Support of innovation

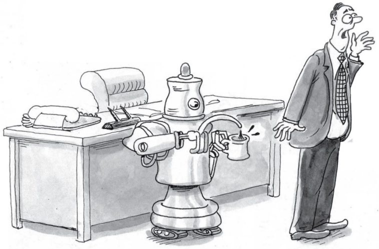
- European Artificial Intelligence Board: cooperation of the national supervisory authorities and the Commission and providing advice and expertise to the Commission; collect and share best practices among the Member States.
- Setting AI regulatory sandboxes
- Facilitating audits of the AI systems
- Establishing a system for registering stand-alone high-risk AI applications in a public EU-wide database
- European common data spaces (e.g. European health data space will facilitate non-discriminatory access to health data and the training on those datasets, in a privacy-preserving, secure, timely, transparent and trustworthy manner)

Compliance with these requirements would imply costs

- For the supply of an average high-risk AI system (EUR 6000–7000)
- For AI users, there would also be the annual cost for the time spent on ensuring human oversight where this is appropriate, depending on the use case (EUR 5000–8000 per year)
- Verification costs (EUR 3000–7500) for suppliers of high-risk AI.

Documenting AI systems

- Relevant documentation and instructions of use; concise and clear information
- Level of accuracy and accuracy metrics should be communicated to the users
- Technical robustnes: risks connected to the limitations of the system (e.g. errors, faults, inconsistencies, unexpected situations) as well as against malicious actions (data poisoning, adversarial attacks)
- Standardisation should play a key role
- Training, validation and testing data sets should be sufficiently relevant, representative and free of errors and complete in view of the intended purpose of the system. They should also have the appropriate statistical properties, including as regards the persons or groups of persons on which the high-risk AI is intended to be used (geographical, behavioural or functional setting)



"Is it harassment if I ask the new guy
to make espresso?"