Fostering Research & Innovation in AI through Regulatory Sandboxes

Lola Montero Santos
Law PhD Researcher
European University Institute
Florence, Italy
lola.monterosantos@eui.eu

Jože M. Rožanec
Artificial Intelligence Laboratory
Jožef Stefan Institute
Ljubljana, Slovenia
joze.rozanec@ijs.si

ABSTRACT

This paper advocates for the establishment of AI regulatory sandboxes in the European Union to enable responsible testing of AI systems in real-life conditions. By aligning the sandbox modalities with the risk tiers of the AI Act, a smooth transition from research to testing of AI systems is ensured. The framework emphasizes the oversight and compliance obligations needed for the desired outcomes to be realised. This will foster AI Research & Innovation in the European Union, delivering benefits for society and ethical legally conforming AI technologies.

KEYWORDS

AI systems, knowledge transfers, EU regulation

1 INTRODUCTION

The European Union (EU) is currently deploying or getting ready to deploy several regulatory instruments to deliver a Union "fit for the digital age"[1]. The not-yet adopted Artificial Intelligence (AI) Act, is one of them. It imposes obligations on providers, makers, and facilitators of AI systems, as well as on users of AI systems or their outputs. The specifics of what constitutes an AI system, the obliged parties, and the conditions these must abide by are still being discussed. The European Commission (EC) released its Proposed AI Act in 2021 [2]. The Council [3] and the European Parliament (EP) [4], have both released their amended versions of the text. These bodies are now engaged in interinstitutional negotiations, which will deliver the Final AI Act, expected by the end of 2023.

The operational functioning of the AI Act will be set at a later stage through implementing acts. However, the content of these documents indicates that regulatory sandboxes will be the chosen environments for the development of safe AI Research & Innovation (R&I). This paper argues that AI regulatory sandboxes should be structured following the tiered approach towards risk that characterises the AI Act, as the space where certain AI systems can be tested before being placed in the market. This framework for AI regulatory sandboxes will favour the growth of AI technologies in the EU and bring about benefits to society.

Information Society 2023, 9–13 October 2023, Ljubljana, Slovenia

2 KEY ASPECTS OF THE AI ACT

To understand the content of this paper, some concepts contained in the AI Act need to be introduced and clarified.

2.1 A Tiered Approach Towards Risk

The Proposed AI Act regulates AI systems based on a tiered approach towards risk. It differentiates between (i) unacceptable risk AI systems, to be outlawed; (ii) high risk AI systems; and (iii) low or minimal risk AI systems. Moreover, the Proposed AI Act sets two categories of high risk AI systems: those characterized by their use as safety components of specific products, and those with implications for fundamental rights. Thus, both the purpose of the AI system and the technologies it utilizes will be key factors in determining the risk category of the AI system. The Final AI Act is expected to follow this structure. However, the specific traits defining what makes the AI systems fall within each category of risk have still not been set. The Final AI Act will likely follow the Proposed AI Act in providing flexibility for the expansion or modification in the future of the traits of AI systems that define them as high risk.

Moreover, the Council and the EP agree with the Proposed AI Act that high risk AI systems will need to be assessed before being put on the market and throughout their lifecycle, while limited-risk AI systems will only need to comply with transparency requirements, enabling users to make informed decisions as to engaging with them. To ease the transition of AI systems from the inception stage to the market stage, the regulation puts forth the creation of AI regulatory sandboxes (sandboxes).

2.2 AI Regulatory Sandboxes

The Proposed AI Act envisions controlled environments for the testing and refinement of AI models, named AI regulatory sandboxes. These are intended to allow obliged parties to ensure that the AI systems comply with the AI Act obligations and to provide feedback on potential risks before such risks can be realized in society. This includes instances of substantial modifications of the AI system which motivates the need for a new conformity assessment. Sandboxes are also intended to enhance legal certainty for AI system innovators.

The concept of regulatory sandboxes is not new. They have been analysed in the literature as experimental regulatory instruments "offer[ing] the flexibility, adaptability, room for compromise, and innovation-friendliness required by novel technological developments" [5]. Regulatory sandboxes have already been implemented across jurisdictions, especially in the

financial sector. They serve companies to test the potential compliance of new business models [6]; and regulators to understand the evolution of new technologies [7] and develop "evidence-based lawmaking" [8].

The Council and EP agree on the creation of AI regulatory sandboxes. Both bodies consider that the specific conditions for the establishment of these environments need to be developed through later delegated implementing acts. Thus, the actual functioning and structure of AI regulatory sandboxes will depend on the implementing acts to be developed and adopted after the Final Text of the AI Act becomes law. The current vision regarding regulatory sandboxes described in the Proposed AI Act and the amendments adopted by the Council and EP contemplates the following stages:

2.2.1 Establishing AI regulatory sandboxes. Specific competent authorities at the Member State(s) and (or) the EU will oversee the accreditation and auditing of these spaces, following given rules and principles. The competent authorities have discretionary powers to adapt their tasks to specific AI sandbox projects.

2.2.2 Conditions of operation of the AI regulatory sandbox. The operation of the AI regulatory sandbox, including the procedure to apply for its utilization, the eligibility criteria, the rights and obligations of participants, duration, and other aspects of operating the AI regulatory sandbox will be set in implementing acts. These sandboxes will be under the direct supervision, guidance, and support of the national competent authority. These are key aspects for the proper functioning and the effectiveness of regulatory sandboxes, as explained by Ranchordas [5].

2.2.3 Modalities of AI regulatory sandboxes. Possibly, different modalities of AI regulatory sandboxes should exist. All sandboxes are intended to deliver controlled environments, permitting the assessment of AI systems before facing full-scale regulatory requirements in real life. The specific requirements and scenarios of different sandboxes are likely to depend on the individual function, technology, or purpose of the given AI systems they are envisioned to assess.

2.2.4 Testing and assessment of AI systems. The sandbox is designed to identify the risks of the AI system, with the purpose of both classifying the AI system accordingly and assuring that the AI system complies with the corresponding rules and obligations. The methods utilized in the AI regulatory sandbox must be geared towards the identification of risks and their mitigation to ensure legal compliance with the AI systems. The AI regulatory sandboxes should focus on dangers to fundamental rights, democracy, the rule of law, health, and the environment. These are, especially, distinguishing traits of high risk AI systems. This way, AI sandboxes can enable truly responsible innovation.

2.2.5 Cooperation among AI Regulatory Sandboxes. The competent authorities should cooperate and coordinate their activities. When possible, cross-border cooperation should be facilitated. This is essential to prevent differences across Member States, and to assure the maintenance of the free movement of products and services in the Union's internal market.

2.2.6 Exclusion of administrative fines by using AI regulatory sandboxes. The sandbox participants that have respected the rules and procedures set within the AI regulatory sandbox framework can enjoy a presumption of legal conformity and will not be subjected to administrative fines for eventual infringements of AI systems legislation, even if they remain liable for the damages they may cause.

In terms of the appropriateness of mainlining the responsibility for potential liability damages during the duration of the sandboxes, the question remains open in the academic sphere. One side agrees with maintaining liability, as the EC and Council defend, arguing that this is necessary for consumer protection and the keeping of trust. However, others consider this approach too onerous, warning that it may disincentivise innovation, and harm smaller players in the market who could be burdened by extensive legal obligations even before fully operating in the market. [9]

2.3 Research Activities & the AI Act

The Proposed AI Act did not include a provision excluding AI research activities from its scope of application. However, both the Council and the EP have brought forth this exemption in their adopted amendments. This suggests that the Final AI Act will set a different framework for such activities.

The Council desires to amend Article 2 of the AI Act to explicitly exclude its application to AI systems "specifically developed and put into service for the sole purpose of scientific research and development", as well as "any research and development activity" [3]. Meanwhile, the EP would amend Article 2 to exclude AI systems research, testing and development activities "prior to this system being placed on the market or put into service" [4]. Neither of these suggested exclusions, however, sufficiently pre-empt potential risks.

This paper argues that for this exemption to operate, the research activity must be performed ensuring the absence of harm to people. Otherwise, research activities that require interaction with people (e.g., to gather behavioural insights, people-facing testing, etc.) could be wrongfully placed outside the scope of the regulation. This could lead to the same societal harms that the AI Act is explicitly tasked to avoid. Thus, this latter type of research activities should also be conducted within the scheme of AI regulatory sandboxes, and their appropriate controlled environment.

3 AI REGULATORY SANDBOXES THAT FOSTER SAFE AI RESEARCH AND INNOVATION

This section argues for the incorporation of three key traits into the framework of AI regulatory sandboxes, either within the AI itself or its delegated implementing acts, for the sandboxes to serve as effective environments for the development of transparent and responsible AI innovation and safe AI systems: (1) making AI regulatory sandboxes the environment for the controlled testing of AI systems in real-life scenarios, (2) creating different modalities of sandboxes following the tiered risk approach of the AI Act and (3) outlining some common requirements for all types of regulatory sandboxes. They also recognize the varying complexities and potential impacts of different AI technologies, ensuring that regulatory oversight is proportionate and targeted to foster the transfer of AI knowledge to society.

3.1 The Shaping of the AI Act Regulatory Sandboxes as the Environment for Real-Life Testing

The Council and EP agree that the 'placing in the market' of the AI system should be the moment when the AI Act is triggered, and the AI system needs to fully comply with the legal obligations within the AI Act. This circumstance is understood as the moment in time in which "[a product] is first supplied for distribution, consumption or use on the market in the course of a commercial activity, whether in return for payment or free of charge" [10]. However, research activities that interact with people in the real world should be covered by AI safeguards, and regulatory sandboxes could provide the entities with means for a progressive transition towards the full applicability of the AI Act.

Currently, the Council and the EP diverge on whether entities should be given the possibility to test AI systems in real-life settings. The Council considers that this should be enabled, under specific conditions and safeguards, within AI regulatory sandboxes. The EP, however, would not exempt the testing of the AI system in real-world conditions from the full application of the AI Act. This paper argues that enabling real-life testing in regulatory sandboxes is the safest and most significant manner in which the AI Act can foster AI R&I while preserving the trust and safety of the people. Real-life testing is necessary. This is in line with the ordinary operation of entities in the market. For example, companies incrementally test whether the changes they implement are successful and behave as expected. If so, they propagate the changes to the rest of their goods or services, while if issues are identified, they revert to the previous version and resolve them.

Carrying out this process for the real-life testing of AI systems within AI regulatory sandboxes, where approval of the AI system is needed before it can be fully released to the market, enables the avoidance of misconduct or abuse. It also ensures that risks are properly identified and mitigated and that by the end of the sandbox period, the outcomes are fully compliant with existing regulations.

3.2 Regulatory Sandboxes Based on the AI's Tiered Approach Towards Risk

This paper argues that AI regulatory sandboxes should be structured following the tiered approach towards risk that characterises the AI Act. Two modalities of regulatory sandboxes can be created according to the potential risk the tested AI systems can generate. These modalities would be foundational, but not exhaustive; others can be created based on criteria such as the sector where the AI system would be deployed.

3.2.1 Regulatory sandboxes for limited-risk AI systems. This sandbox would serve to test new limited-risk AI systems, or those which are already in the market, but are being applied to an additional or different purpose. Access to such a sandbox should be voluntary, and legal requirements less strict.

3.2.2 Regulatory sandboxes for (potentially) high risk AI systems. This sandbox would test new high risk applications, or existing high risk AI systems for a new purpose. This sandbox should also be utilised if the entity is unsure about the risk classification of the AI system. The main purposes of this modality are to enable entities to (1) test their AI system, to assess whether it is high risk, and (2) if the AI system is high risk, to determine what mitigating factors can be implemented, and if the implemented mitigated factors are sufficient. The utilisation of this type of sandbox could be voluntary or compulsory. The choice depends on the ability of certification bodies to establish sufficient high risk AI systems regulatory sandboxes, and the associated benefits the entities utilising them could enjoy. Making the utilisation of this sandbox compulsory is the most effective way of assuring that high risk AI systems conform to the law before being placed in the market. If the utilisation of this sandbox is made voluntary, its use could provide the entity with a fast-tracking process in the third-party conformity assessment procedure all high risk AI systems must undergo.

Moreover, certain entities utilising this type of sandbox could be given access to a 'nursery status', a concept developed in other jurisdictions. This status acts as a transitional phase where companies, especially startups, can continue to receive targeted support even after exiting the sandbox environment. This responds to the fact that startups often rely heavily on the guidance provided during the sandbox period, unlike established companies that are more experienced in the field of regulatory compliance. The nursery status recognizes that, mitigating the risks of no longer being exempt from regulatory consequences, and facing real-world responsibilities (including potential fines), by offering increased support. This continued assistance helps organizations meet regulatory requirements and build the necessary experience in a more controlled setting, serving as a period of growth. [11]

3.3 Common requirements for all Regulatory Sandboxes

Regulatory sandboxes must adhere to certain common requirements to ensure that AI systems and other innovative technologies go through real-life testing within controlled and legally compliant environments. These minimum terms and conditions must be explicitly defined, as part of the procedure to establish the regulatory sandbox. The requirements for limited-minimal risk AI sandboxes can be adjusted, reflecting the lower danger posed by such AI systems. This section argues that all AI regulatory sandboxes must meet the following criteria:

3.3.1 The identification of the AI system features that are being tested. This encompasses understanding not only what functionalities are being tested but also why and how they are being assessed. The supervisory authority will not have direct access to the code itself and must safeguard sensitive and(or) proprietary information, allowing innovation to flourish without undue risk of exposure.

3.3.2 The proportion, composition, and selection of users subjected to testing. Users should be made aware that they are engaging with an AI system that is being tested, and must provide their consent. For instance, if a financial institution is offering a new credit product based on an experimental algorithm, customers must be informed that this offering is not part of the financial institution's regular operation.

3.3.3 The time frame for testing, with provisions to interrupt it. The complexity of the technology and the nature of the testing environment should justify the start and end dates of the

regulatory sandbox. Crucially, provisions must be made to allow for an immediate interruption of the testing if insurmountable risks arise, with an identification of the measures set to identify such a situation.

3.3.4. Documentation and timestamping. Entities benefiting from regulatory sandboxes must develop rigorous documentation. This may include timestamps indicating when specific documents, descriptions, or test plans were submitted. As a counterpart, entities could utilise this document to undergo or strengthen their claims over intellectual property rights.

4 BENEFITS OF REGULATORY SANDBOXES

Regulatory sandboxes can be constituted as the best environment to achieve legally conforming AI systems being released to the market. They entail benefits for the various stakeholders:

4.1 AI System Innovator

The AI regulatory sandbox enables the testing of new technologies that do not yet exist in the market and may therefore still not be subjected to a given classification, or which need to be modified to mitigate risks. In cases where the use of the AI regulatory sandbox has not served to prevent the materialisation of risk, the company utilizing the AI system may still be considered liable for the harms incurred, but the companies will not be fined for unexpected harms of the AI system.

The UK experience with regulatory sandboxes reveals other associated benefits. Among them, sandboxes have been found to improve access to capital, as firms operating within these controlled environments often find it easier to secure investment. These firms are also more likely to remain in operation and even secure a patent. Sandboxes also significantly reduce the time and cost of getting products to market, a factor that is particularly beneficial for first-time innovators. [12]

4.2 AI System Regulators

The regulatory sandboxes permit the establishment of feedback loops in the regulation. Regulators themselves can observe if the sandboxes are meeting their desired goals, or whether some AI systems need to transit from one category of risk to another. In cases of AI systems causing harm despite being considered legally compliant by AI regulatory sandboxes, the regulators can update the functioning of the AI regulatory sandboxes, to avoid this from happening again.

4.3 Benefits for Society at Large

The purpose of the AI Act is to foster safe innovation. Regulatory sandboxes would enable this, but also an increased degree of positive spillover effects for society. The sandbox, by improving the collaboration between the regulator and the innovator, has the potential to enhance consumer protection by fostering a more transparent and cooperative relationship that focuses on safety and compliance. Another significant benefit is the increased throughput of tested and introduced products and services to the market. Regulatory uncertainty frequently inhibits the most innovative products from reaching consumers, as they are often abandoned at early stages due to associated risks. Through the sandbox framework, these products can be guided and supported, thereby minimizing early-stage abandonment and enhancing the flow of innovative solutions into the marketplace.

5 CONCLUSION

This paper contends that AI regulatory sandboxes must be established as the natural environment for the controlled testing of AI systems within the EU. By aligning sandboxes with the tiered risk approach of the AI Act, two main modalities of AI Regulatory Sandboxes can be created, tailored to the potential limited-minimal risk, or high-level risk of the AI system. This structure not only facilitates a seamless transition from research to testing but also ensures strict, transparent oversight of AI technologies. By integrating provisions for user consent, intellectual property protection, defined time frames, and safeguards against risks, these measures will propel the growth of AI technologies in the Union, while allowing the systematic and informed integration of AI technologies into broader societal contexts and applications.

ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency and the European Union's Horizon Europe research and innovation program project Graph-Massivizer under grant agreement HE-101093202.

REFERENCES

[1] von der Leyen U, 'Political Guidelines for the next European Commission 2019-2024' (2019) 13.

[2] European Commission, Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts 2021 [COM(2021) 206 final]. Released on 21 April 2021.

[3] Amendments of the Council of the European Union on the Proposed AI Act. Interinstitutional File: 2021/0106(COD) 14954/22. Adopted on 25 November 2022.

[4] Amendments of the European Parliament on the Proposed AI Act. File: P9_TA(2023)0236. Adopted on 14 June 2023.

[5] Ranchordas S, 'Experimental Lawmaking in the EU: Regulatory Sandboxes' (2021) <https://papers.ssrn.com/abstract=3963810>

[6] 'Regulatory Sandbox' (FCA, 1 March 2022) <https://www.fca.org.uk/firms/innovation/regulatory-sandbox>

[7] Ahern DM, 'Regulatory Lag, Regulatory Friction and Regulatory Transition as FinTech Disenablers: Calibrating an EU Response to the Regulatory Sandbox Phenomenon' (22 September 2021) <https://papers.ssrn.com/abstract=3928615>

[8] Pop F and Adomavicius L, 'Sandboxes for Responsible Artificial Intelligence' (European Institute of Public Administration 2021) Briefing <https://www.eipa.eu/publications/briefing/sandboxes-for-responsible-artificial-intelligence/>

[9] Truby J, Brown RD, Ibrahim IA and Parellada OC, 'A Sandbox Approach to Regulating High-Risk Artificial Intelligence Applications' (2022) 13 European Journal of Risk Regulation 270 <https://www.cambridge.org/core/journals/european-journal-of-risk-regulation/article/sandbox-approach-to-regulating-highrisk-artificial-intelligence-applications/C350EADFB379465E7F4A95B973A4977D#fn18>

[10] UK notice 'Definition of 'placing on the market' before and after the UK leaves the EU, if there's no Brexit deal' <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/792539/placing-on-market-definition.pdf>

[11] Johnson WG, 'Caught in Quicksand? Compliance and Legitimacy Challenges in Using Regulatory Sandboxes to Manage Emerging Technologies' (2023) 17 Regulation & Governance 709 <https://onlinelibrary.wiley.com/doi/abs/10.1111/rego.12487>

[12] Cornelli G, Doerr S, Gambacorta L and Merrouche O, 'BIS Working Papers No 901 Regulatory sandboxes and fintech funding: evidence from the UK'. (November 2020, revised April 2023) Monetary and Economic Department of the Bank for International Settlements <https://www.bis.org/publ/work901.pdf>