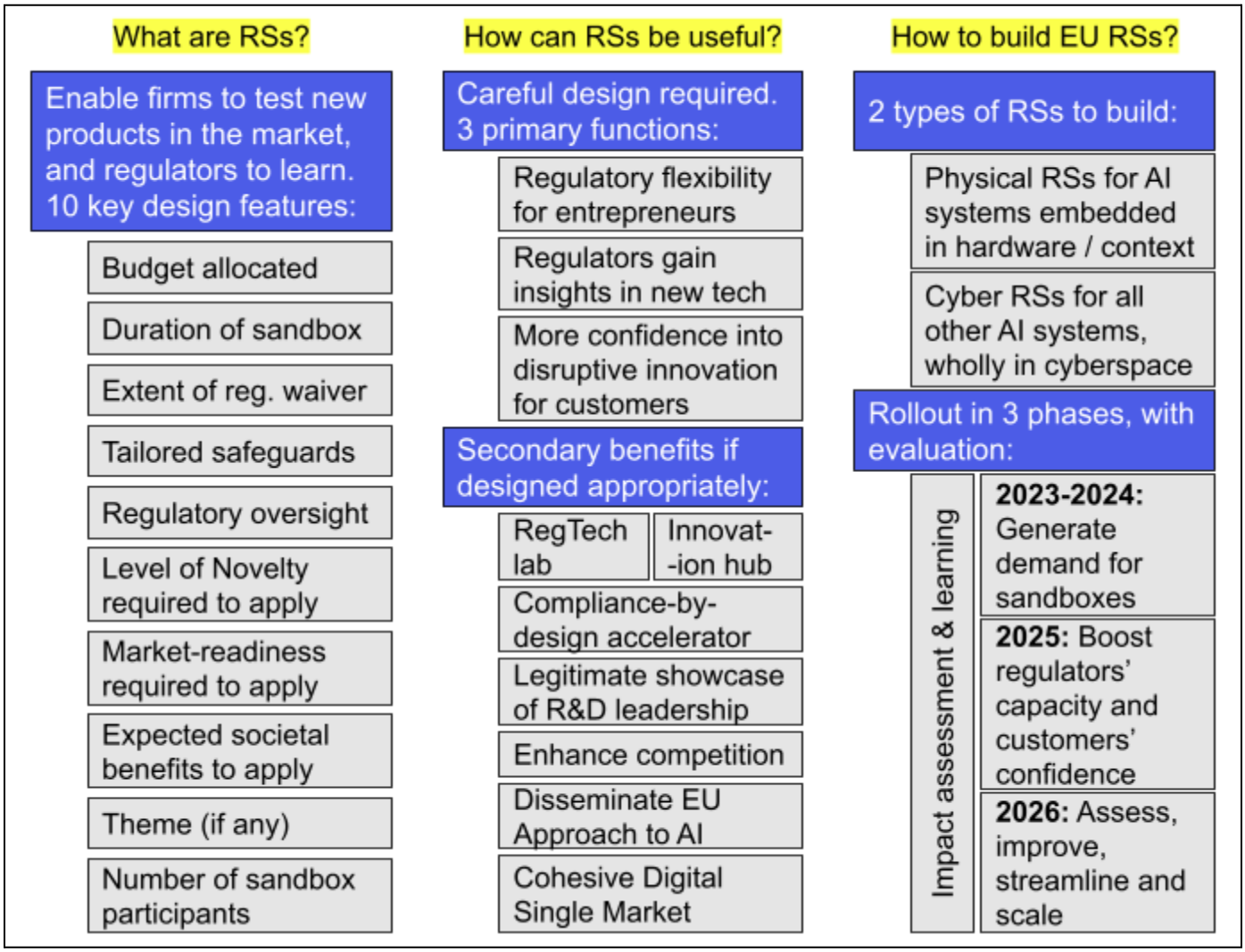
THE FUTURE SOCIETY

Sandboxes without the quicksand: making EU AI sandboxing work for regulators, entrepreneurs and society

Memo by The Future Society (TFS),[1] February 2022

The EU AI Act has introduced the concept of **AI regulatory sandboxes** (henceforth **'RSs'**) as a measure in support of innovation.[2] RSs are particularly sophisticated governance mechanisms that can reconcile the objectives and interests of regulators, entrepreneurs and society at large, but whose success and cost-effectiveness require careful design. In this memo, we briefly explain what RSs are, their functions, and our recommendations to adapt them for the EU.



[1] Contact: Nicolas Moës, Head of Operations (nicolas.moes@thefuturesociety.org, +32 488 541917)
[2] Proposal for a Regulation on a European approach for Artificial Intelligence, Article 53 'AI regulatory sandboxes'

What are Regulatory Sandboxes?

Regulatory Sandboxes (RSs) are time-bound programmes administered by authorities. They enable entrepreneurs to test new products and services in the market, but with increased oversight from authorities. Some RSs reduce regulatory responsibilities for entrepreneurs, replaced by tailor-made safeguards for society enforced by regulators.[3][4][5] To clear any confusion between policymakers and software developers in current AI Act conversations: Regulatory Sandboxes are *not* software sandboxes.[6] Over the past 15 years, RSs have been used across a broad range of sectors undergoing digitalization, including finance, energy, health, transports, and legal services.[7]

The typical process for a RS involves a sectoral authority or issue-specific agency to launch a public call for applications to enter the sandbox. It will collect and assess applicants (firms that have a sandbox-ready product) and select those most promising based on pre-established criteria. The selected applicants ("sandbox participants") will then begin a time-bound product testing programme in the market under the supervision of the authority. After the time is up, the product will exit the RS, either withdrawn from market for further rework until demonstrably compliant with the regulations in place, or launched publicly on the market and subject to the regulations in force.

Sandboxes have 10 key design features to calibrate to the authority's objectives:

1. The **resources allocated** and whether the authority staff operates in a standalone, dedicated department or agency or in a taskforce format determine the overall burden that the sandboxes represent for the civil service. Traditional regulatory sandboxes require significant staff time to manage successfully.
2. The **time limit** determines how long the AI system can be tested in the sandbox, ranging from 2 weeks to 2 years, typically between 6 and 12 months. Not all AI systems' live testing require the same amount of time, so each sandbox could have a different time limit. Nevertheless, it is important to time-bound the programme to avoid moribond innovations to consume public resources for minimal societal benefits.
3. The **extent of the regulatory waiver**, if any, determines which post-deployment requirements and obligations the sandbox participants can temporarily be exempted from. These exemptions are strictly-controlled relaxation of the rule of law, in direct collaboration with authorities, for the entities participating in the sandbox.
4. The **safeguard mechanisms** to offset this waiver are sandbox-specific measures jointly developed and enforced by authorities and entrepreneurs to achieve the regulatory

[3] Council of the European Union (2020) "Outcomes of Proceedings 13026/20 BETREG 27"

[4] *Omarova, S.T. (2020), "Dealing with Disruption: Emerging Approaches to Fintech Regulation", Washington University Journal of Law & Policy* 61(1): 25-54

[5] World Bank Group (2020), "Global Experiences from Regulatory Sandboxes" Finance, Competitiveness & Innovation Global Practice, Fintech Note (8)

[6] Software sandboxes are virtual environments isolated from deployment servers and used for testing software code without affecting the deployed version of the software. Software sandboxes are also known as test servers or development servers.

[7] Attrey, A., Lesher, M. and Lomax, C., 'The role of sandboxes in promoting flexibility and innovation in the digital age', OECD Going Digital Toolkit Policy Note 2, 2020

objectives (protection of health, safety and fundamental rights) while providing some procedural or technological flexibility in the achievement of these objectives.

5. The **intensity of regulatory oversight** that helps ensure these safeguard mechanisms are properly designed and enforced. This oversight is naturally accompanied by more or less intensive learning about new technologies that challenge current regulations, which informs regulatory adjustments. This intensity is proportional to the resources allocated to the sandbox by the authorities.
6. The **level of novelty required** to justify entering a sandbox is an important selection criterion, combined with justification of why this novelty cannot flourish under the existing regulatory environment. If this level is too low, the sandbox is open to any company that creates marginally innovative products. This undermines the rule of law: the baseline regulation waived in the sandboxes applies only to old products and probably needs to be updated. If this level is too high, the sandbox is only accessible to once-in-a-decade disruptive innovations, which might hinder innovations that could create new markets if given the opportunity but that are nevertheless preempted from the existing baseline regulation.
7. The **market-readiness required** to be eligible to enter the sandbox affects its overall impact. At one extreme, sandboxes could require “pre-registration” of AI R&D projects to develop a landscape overview, but then it would have to funnel these down over time to invest resources only on these projects that make it to market. At the other extreme, the sandbox could require market-ready technologies that are about to pass through the conformity assessment process, but then the sandbox cannot enable compliance-by-design.
8. The **expected consumer and societal benefits,** in balance with the potential consumer and societal harms that the innovation could generate, help the authorities assess whether the sandboxing is worth the opportunity cost of the public resources dedicated to the sandbox.
9. The **specific theme** of the sandbox determines what type of innovation authorities are ready to create sandboxes around, for what sectors or functions, and potentially, in which geographies. In the context of the AI Act, the type of innovation is AI; however, it might be at times necessary to specify a sector or function targeted by the RS, to ensure that the relevant authorities are involved institutionally in sandbox oversight.
10. The **number of participants** and the selection process determine how many innovations can be live-tested in parallel (in multiple sandboxes within the same sandbox programme). Once the processes are set up and the fixed costs incurred, it may make budgetary sense to increase the capacity of the sandbox programme, though there are still some variable costs that would increase proportionately with the number of participants.

How can RSs be useful?

RSs effectiveness for regulators, entrepreneurs and society at large depends on very careful design, engineering, and management.[8] For example, undue influence on the selection process, lack of safeguards or miscalibrated scope could undermine the rule of law or create anti-competitive distortions. RSs are also very resource-intensive, so return on investment must be carefully assessed.

However, if they are properly designed and managed, RSs can fulfill multiple primary and secondary functions. The primary functions they should fulfill are:

- **Providing monitored regulatory flexibility for entrepreneurs to "live" test their innovations on the market.** Instead of going into "stealth mode" and attempting to pass under the authorities' radar until they are profitable, innovators can work transparently and hand-in-hand with regulators for compliance-by-design.
- **Providing regulators with a greater understanding of the cutting-edge technologies, value chains, and business models.** These insights can help them assess the fitness of the regulatory environment and suggest improvements to policymakers.
- **Providing customers (B2B or B2C) with the confidence to do business with disruptive technology providers.** Users, and to a certain extent the rest of society, will benefit from adopting innovative products whilst knowing their fundamental rights, health and safety are protected. They also benefit from more competitive markets with lower barriers to entry, disrupting existing players.

Depending on its design, its ambitions and the resources allocated to it, the EU AI Act sandbox system could also fulfill additional, secondary functions for society. These include:

- Serving as a central, one-stop solution for entrepreneurs to address queries related to regulations, administration, and policies ("innovation hub"). Indeed, one of the most robust empirical findings on sandboxes so far is that greater communication between regulators and entrepreneurs is key for better market governance.[9]
- Giving regulators the opportunity to experiment with their own regulatory tools and protocols to gather empirical evidence with the goal of improving enforcement or informing policy development ("RegTech labs").
- Facilitating compliance-by-design by making end-to-end support available to all sandbox candidates, allowing them to navigate the best practices and harmonized standards for their regulatory requirements and obligations ("compliance-by-design accelerator").
- For early stage products, offering a form of officially recognised pre-registration of innovative R&D projects, showcasing thought-leadership in the field and, upon graduation from the sandboxes, high-quality products that are compliant by design.

[8] Ranchordas, S. (2021) Experimental Regulations for AI: Sandboxes for Morals and Mores
[9] Buckley, R.P. et al. (2020) Building Fintech Ecosystems: Regulatory Sandboxes, Innovation Hubs and Beyond

- Lowering barriers to entry into markets by de-risking market experimentation and reducing time to market by clearing regulatory hurdles without creating competitive distortion vis-a-vis the candidates not selected for the sandboxes.
- Helping disseminate the European Approach to AI by attracting entrepreneurs from all over the world to apply to the sandboxes and become closely familiar with their rules.
- Supporting a cohesive Digital Single Market by making an EU-wide sandbox system and avoiding cross-border regulatory arbitrage.

How to build EU AI RSs?

Launch two types of RSs: Physical RSs and Cyber RSs

Given the level of care and amount of resources required for developing cost-effective RSs and the many additional benefits they could provide, **we recommend developing two types of RSs corresponding to two different types of AI systems.** This distinction follows the EU AI Act's separation between AI systems embedded in products and services and stand-alone AI systems.[10]

1. **Physical RSs** for AI systems embedded in physical products or services. These AI systems are functionally dependent on hardware, environmental input or the physical context for their proper operation and testing. These include for example: autonomous vehicles, smart advertising signs, AI-enabled supermarket cash registers, AI-enabled warehousing, smart medical devices, etc.
2. **Cyber RSs** for AI systems operated and used on a stand-alone basis, not embedded in physical products or services. These AI systems are developed and deployed in cyberspace and whose proper operation and testing do not depend functionally on environmental input and physical context other than the user's device. These include for example: chatbots, coding assistants, general purpose AI systems, recommender algorithms, analytics algorithms, AI systems being trained in virtual environments and simulations, etc..

The distinction between Physical and Cyber RSs is important in terms of cost-effectiveness because we expect Cyber RSs to be much less costly to roll out. Cyber RSs would rely mostly on software tools, shared virtual environments and online communications, very similar to how software alpha and beta testing[11] take place today. In comparison, **Physical RSs would rely on on-site demonstrations and monitored trial periods in the field**, which would require identifying a physical site in which to deploy or embed the AI system and coordinating with hardware providers, authorities and local stakeholders. We expect this to take significantly more time than with cyberspace stakeholders, given there exist

[10] Page 18, paragraph (6) in Proposal for a Regulation on a European approach for Artificial Intelligence
[11] Alpha and beta testing are phases in the software release lifecycle. Alpha testing corresponds to an assessment of the software before inclusion of all its functional features and before making it accessible to users. Beta testing corresponds to testing of software when it contains all its functional features. Software in the beta testing phase is sometimes made available to users in order to assess its usability.

fewer regulations governing cyberspace. Moreover, tools and safeguard mechanisms for one Cyber RS could transfer to many Cyber RSs with minor tailoring, while safeguard mechanisms for Physical RSs would have to be tailored Note that an AI system can pass through both sandboxes: for example, a general-purpose computer vision system can first be tested in a Cyber RS for its general-purpose abilities and then in a Physical RS when it is integrated into an autonomous vehicle.

Rolling out sandboxes: phasing-in the various benefits

As described above, a sandboxing system can aim to fulfill many functions. Given the complexity and the level of care needed to successfully design cost-effective RSs, **we recommend developing the sandboxing ecosystem in three phases, with each phase building upon the previous phase's successes and lessons learned.**

Phase 1: Generate demand for sandboxes (2023-2024)
The purpose of this phase is to roll out the basic features of the EU sandboxing ecosystem in order to generate demand for sandboxes from both entrepreneurs and regulators. To do so, it is important to capture the low hanging fruits that RSs can provide. These low hanging fruits include:

- Enabling entrepreneurs to live test their innovations on the market under monitored regulatory flexibility.
- Enabling regulators to obtain a greater understanding of the cutting-edge technologies and business models.
- Making the Digital Single Market more cohesive.
- Spreading the European Approach to AI abroad.

Achieving this will require the creation of an "innovation hub" and the acceleration of compliance-by-design. An innovation hub is "*a dedicated point of contact for firms to raise enquiries with competent authorities [...] and to seek non-binding guidance on the conformity of innovative [products, services or business models]*".[12] It is open to all entrepreneurs, not only to sandbox participants. It enables regulators and entrepreneurs to develop a mutual understanding and connections, which is fundamental to the proper functioning of a sandboxing ecosystem. Accelerating compliance-by-design involves providing entrepreneurs with the tools to comply early on in the development of their product (e.g. actionable best practices and harmonized standards to comply with regulatory requirements and obligations, calibrated testing environments and protocols, guidance on the latest evolution of AI-related compliance, etc.)

To avoid fragmentation of the Digital Single Market, the sandboxing system would be centralized at the EU level, either in a standalone EU agency/department for AI or within DG Connect. Close cooperation with national authorities to ensure best practices and lessons learned are

[12] Joint Committee of European Supervisory Authorities (2018) FinTech: Regulatory sandboxes and innovation hubs

shared throughout the EU and to avoid jurisdictional conflicts.[13] This need for cooperation is particularly acute for Physical RSs, where local authorities will need to be involved as well. To ensure that the RSs attract entrepreneurs from all over the world and to avoid trade barriers, they should be open to foreign entities ready to develop “Tech fit for EU”.

Moreover, given the additional regulatory scrutiny, generating demand for sandboxes from entrepreneurs will require including several benefits. These benefits need not be expensive to taxpayers. Some already exist and could be re-used from other EU programmes. This could include granting access to pre-deployment services, such as assistance for preliminary registration of AI systems, compliance and R&D support services; facilitating contacts with all the other relevant elements of the Union’s AI ecosystem and other Digital Single Market initiatives such as Testing & Experimentation Facilities, Digital Hubs, Centers of Excellence; and granting access to standardization documents and certification, community social platform and contact databases, the existing portal for EU tenders & grantmaking, and to a list of investors.

Finally, the first phase should also build-in impact assessment and documentation processes to facilitate evaluation of the sandboxing ecosystem’s effectiveness. Building RSs is very difficult, and the EU authorities will have to experiment and learn throughout the rollout to optimize their approach and operations and avoid failure.

Phase 2: Boost regulators’ capacity & foresight and customers’ confidence (2025)
The purpose of this phase is to fully tap the value that a sandboxing system can generate for regulators on one hand and society at large on the other hand. To do so, phase 2 requires investing in the “institutional infrastructure” surrounding the RSs.

For customers and society at large, phase 2 must provide clear and accessible information about AI systems, in particular systems exiting RSs to ensure evidence-based trust in novel AI systems. This information can range from more pedagogical technical documentation to a labeling, scoring or metrology system to convey relevant information about AI systems, based on robust audits and other assessments. Enabling entrepreneurs’ public pre-registration in RSs would also enable them to showcase their thought-leadership to customers, who could then more easily plan purchases at the cutting edge of the current technology frontier.

For regulators, phase 2 focuses on ensuring that they have proper tools and competence for safeguarding customers of sandbox participants and for learning. For example, this could involve the development of a free and confidential compute-intensive auditing service to help understand the "real world” behavior of certain very promising but opaque AI systems, such as general purpose AI systems. To develop and continuously update this regulatory toolkit, the sandboxing ecosystem should comprise a RegTech lab: an internal programme enabling regulators to experiment with their own regulatory tools and protocols (e.g. conformity assessment) within the RSs and to gather empirical evidence to better inform enforcement and

[13] Allen, H (2019), “Regulatory sandboxes”, *The George Washington Law Review* 87(3): 580-643.

policy development. The lab would aim to future-proof regulatory capabilities: enabling the foresight needed today to govern tomorrow's innovations.

For both customers and regulators, phase 2 will require investing in the institutional ecosystem surrounding RSs, in particular in cost-effective means to measure, assess and communicate that assessment about relevant dimensions of the AI systems (accuracy, robustness, cybersecurity, transparency, etc.). Cyber RSs in particular should include a suite of software tools that entrepreneurs and regulators can leverage to jointly test and assess their systems prior to launch. Given the significant research and engineering work required for developing and deploying these capabilities, it is important to leverage existing workstreams. For example, while the EU-level capabilities to benchmark relevant dimensions of AI systems is currently limited, the European Commission is currently developing Testing & Experimentation Facilities which could contribute to building these capabilities.

Phase 3: Assess, improve, streamline and scale (2026)
The purpose of this phase is to evaluate and improve the ecosystem based on the information generated by impact assessments and documentation processes and on the experience of individuals involved. This major review should determine whether the ecosystem is fulfilling its objectives and, if not, whether it should be sunset or redesigned for addressing its shortcomings. This review will help determine what budget to allocate structurally to the sandboxing system.

In this phase, the authorities in charge of running the sandboxing system will develop plans for streamlining the regular opening of sandboxes -informed by the review and lessons learned- and for scaling the system to broaden its impact. The distinction between Cyber and Physical RSs is crucial for scaling and streamlining: while it seems operationally difficult to run any more than 30 Physical RSs at any given time (each with multiple participants), Cyber RSs are organisationally much cheaper and could reach many times that number, depending on the extent of regulatory oversight needed. It is premature to try and establish these figures ex ante, or to try to determine exactly how the sandboxing system will look like "at scale": a lot of it should and will depend on the experience accumulated over the previous 4 years.