UNIVERSITÀ DEGLI STUDI FIRENZE
Dipartimento di Scienze Giuridiche
Eccellenza 2023-2027

CYBER RIGHTS

POLICY NOTE #01

Sandboxing Governance Models for Trustworthy AI and Data Use

September 2026

This initiative was supported by the EU-funded EUSAiR project.

EUSAiR
BRIDGING INNOVATION AND COMPLIANCE IN AI

Note: The image has been artificially generated.

Context:

This policy note synthesises insights collected through multiple complementary sources: discussions held during a workshop at the Computers, Privacy and Data Protection (CPDP) conference (Brussels, 20 May 2026), written contributions submitted by workshop participants following the event, and targeted expert input gathered from additional stakeholders.

The workshop was an initiative of the CybeRights Center and was organised by **Andrea Simoncini** (Professor of Constitutional Law, University of Florence), **Erik Longo** (Professor of Constitutional Law, University of Florence), **Filippo Bagni** (Research Fellow at CybeRights Center) and **Fabio Seferi** (PhD Candidate in Cybersecurity, IMT School for Advanced Studies Lucca and University of Florence), with the operational support of **Emma Bagnulo** (PhD Candidate in Cybersecurity, IMT School for Advanced Studies Lucca and University of Florence). The initiative was supported by the project "EU Regulatory Sandboxes for AI" (EUSAiR) – DIGITAL-2024-AI-ACT-06-SANDBOX.

Contributing workshop participants reflect a broad cross-section of academia, technical research, public policy, and industry practice: **Arvin Khozooei** (Owner of Aethico and AI Governance Lead at TU Delft); **Dave Lewis** (Professor in Computer Science, School of Computer Science and Statistics); **Thiago Guimarães Moraes** (PhD Researcher, at the Vrije Universiteit Brussel - Law, Science, Technology and Society research group (VUB-LSTS) and University of Brasília (UnB)); **Juliana Oms** (Expert Advisor on digital public policy at the Brazilian Internet Steering Committee (CGI.br)); **Emanuele Parisini** (Policy Officer AI Unit, EDPS); **Antonino Rotolo** (Head of Alma AI, University of Bologna; Coordinator of EUSAiR Project); **Clément Sastre Barbié**; **Thomas Streinz** (Joint Chair in Law and Regulatory Theory at the European University Institute); **Sophie Weerts** (Professor of Public Law and Regulation, IDHEAP – Institut de hautes études en administration publique, Swiss Graduate School of Public Administration, Université de Lausanne).

Additional expert contributions were provided by stakeholders across legal practice, research, international governance networks, and public administration: **Davide Baldini** (PhD Candidate in European and Transnational Legal Studies, University of Florence and Maastricht University; Managing Partner, ICTLC - ICT Legal Consulting); **Andrea Carrubba** (Public Servant, Tuscany Region); **Armando Guio Español** (Research Affiliate, Harvard Law School; Executive Director, Global Network of Internet & Society Centers (NoC)); **Imane Hmiddou** (Ing. Mgr. MA; Doctoral Researcher at University of Bologna; EUSAiR Team); **Devit Keqi** (PhD Candidate University of Florence - Public Servant, Tuscany Region); **Sophie Tomlinson** (Deputy Executive Director, Datasphere Initiative); **Gianluca Vannuccini** (Director of SIITI Directorate, Tuscany Region); **Raphael von Thiessen** (Programme Lead AI, Office for Economy, Canton of Zurich); **Laura Zoboli** (Assistant Professor, IE University – School of Law; Lead for Regulatory Sandbox activities, VIPPSTAR (Horizon Europe Project 101156763)).

Note: The list of acknowledged participants and contributors includes only those who consented to formal acknowledgment.

The policy note was edited by **Fabio Seferi**.

Citation: CybeRights Center (2026), *Sandboxing Governance Models for Trustworthy AI and Data Use*, Policy note No. 1, https://www.cyberights.unifi.it/.

Table of Contents

0. Executive summary

AI regulatory sandboxes[1] are among the most closely watched governance instruments in the AI Act. Across all three channels, a consistent message emerges: sandboxes will only fulfil their regulatory promise if competent authorities deliberately design them as learning infrastructures rather than as compliance shortcuts for individual providers.

Contributors converge on four connected priorities. First, sandboxes need an explicit mandate, resources and institutional willingness to generate transferable regulatory knowledge, not only to help individual participants reach the market. Second, the multiplicity of authorities with a stake in any given AI system – data protection authorities, market surveillance authorities, the AI Office, the European Data Protection Supervisor (EDPS), and sectoral regulators – requires structured coordination mechanisms decided at the design stage, rather than ad hoc involvement negotiated case by case. Third, sandboxes should be expected to produce a wide range of evidence, including failure cases and organisational and rights-impact evidence, and this evidence should inform better implementation and guidance rather than automatically justify weaker safeguards. Fourth, governance design choices – who gets access, how transparent the process is, how affected communities are involved, and how trade secrets are protected – are not secondary implementation details but the substance of what makes a sandbox trustworthy.

The note closes with a consolidated set of recommendations for competent authorities that are establishing or operating regulatory sandboxes, organised around various themes: institutional design and mandate; multi-level coordination; evidence generation and use; transparency and stakeholder participation. EU-level learning and dissemination; and standardising regulatory learning across fora beyond the sandbox itself.

1. Introduction and methodology

This note draws on three complementary sources of input, all structured around the same four guiding questions concerning AI regulatory sandboxes under the AI Act:

1. Beyond compliance: what would it take for AI regulatory sandboxes under the AI Act to operate as genuine laboratories for re-thinking data governance and accountability, rather than remaining narrow compliance tools for individual providers?
2. How should responsibilities, powers and expectations be redistributed between EU-level bodies (such as the AI Office and EDPS), national data protection and market surveillance authorities, and sandbox participants, so that experimentation strengthens accountability in AI and data protection?
3. What forms of evidence and learning (technical benchmarks, organisational practices, rights-impact narratives, failure cases, citizen input) should sandboxes be expected to generate, and which types of evidence should not be used to justify regulatory relaxation or "simplification" in the name of innovation?
4. Which concrete design choices for sandbox governance (access criteria, transparency duties, participation of affected communities, safeguards for trade secrets and fundamental rights) best

[1] In the present policy note, the terms "AI regulatory sandbox", "regulatory sandbox", and "sandbox" are used interchangeably.

support trustworthy AI, and how can their outcomes be documented and shared to inform the next generation of EU digital regulation inside and beyond Brussels?

The first source of input is the **live discussion captured during the CPDP workshop** itself, reflecting an exchange among practitioners, regulators, and researchers. The second consists of **structured written contributions submitted by workshop participants after the event**, focused in particular on the fourth question, which was not sufficiently tackled during the workshop. The third consists of **independent written contributions submitted by experts** who were unable to attend the workshop, but who responded to all four questions. Taken together, these sources reflect the perspectives of legal academics, sandbox operators, industry participants and civil-society-oriented contributors. This note distils recurring themes and points of tension across all three sources; it does not attribute specific positions to specific individuals, and it does not purport to represent a single institutional view.

2. From compliance instrument to learning laboratory

Nearly every contribution warns against a narrow reading of sandboxes as ex-ante compliance-support mechanisms whose sole purpose is to help an individual provider demonstrate conformity with a pre-defined set of legal requirements. That function is real and valuable, particularly for smaller organisations that lack in-house regulatory expertise, but contributors are emphatic that it should not exhaust the ambition of the instrument. One contribution pushes back on the "laboratory" framing itself: a sandbox is not a scientific experiment, since it typically lacks a research question, a hypothesis, or a methodology capable of guaranteeing reproducible results. On this view, the priority should not be to make sandboxes more "scientific" in form, but to secure three concrete conditions in practice: (i) project submissions should clearly identify the specific data-governance or accountability issue the project is meant to illuminate; (ii) the public officials running the sandbox should have adequate legal and technical expertise, sufficient time and genuine independence to investigate the project, supported where possible by an internal peer-review mechanism to limit individual bias; and (iii) there should be a real mechanism for lessons to reach EU-level policymakers who are prepared to act on them – since even the most sophisticated governance design will fail to achieve its purpose if policymakers do not, in the end, listen.

2.1 Sandboxes as feedback loops into the regulatory framework

A recurring idea is that sandboxes should create a **structured feedback loop** into the application and, where necessary, the evolution of the AI Act itself. As a new and rapidly-tested regulatory framework, the AI Act contains open-textured obligations – for example on human oversight or on what makes training data "representative" or "free of errors" – whose operational meaning can only be established through practical experience. Contributors argue that without this feedback loop into rule-making and guidance, the systemic impact of sandboxes will remain limited, however useful they may be to individual participants. Sandboxes also need mechanisms to keep pace with fast-moving technology, particularly agentic and general-purpose systems, so that regulatory learning addresses not only current requirements but also emerging risks. This feedback function should also be applied to regional or local sandboxes. Where a sandbox is run by a regional or local authority, the obligations of the AI Act are tested against concrete, place-specific administrative practices; the resulting learning should feed not only upward into EU rule-making, but also horizontally, across comparable regions facing similar operational conditions, and downward into the design of local digital public services.

2.2 Testing the operability of the regulatory framework, not only compliance

Several contributions distinguish between testing whether a specific AI system complies with the law and testing whether the legal requirements themselves can be meaningfully operationalised when several regulatory regimes apply simultaneously. In highly regulated sectors, an AI system may fall under the AI Act, the GDPR, the Cyber Resilience Act, the Medical Device Regulation, the Platform Work Directive, or the Digital Services Act at the same time, and the central difficulty is often not whether a provider understands any single regime, but whether the regimes function coherently when confronted with a real system. One contribution points to a national trilateral pilot combining telecommunications, data protection and innovation authorities as an illustration: its principal value lay not in clearing individual use cases but in revealing where regulatory frameworks overlap or collide in practice. On this reading, sandboxes are uniquely positioned to **assess operability across regimes**, a function that ordinary compliance instruments cannot perform.

2.3 Data governance and the special-data provisions of the AI Act

Robust, representative and fair datasets remain scarce in several sectors, which slows both the trustworthiness and the market development of AI in the EU. Contributors see an opportunity for sandboxes to **support structured collaboration** between data spaces, the authorities responsible for the Data Governance Act, national statistical bodies, data protection authorities, Testing and Experimentation Facilities (TEFs) and relevant research institutes, helping competent authorities map available datasets, identify gaps and prioritise investment.

Particular attention is paid to the exceptional **data-processing possibilities in the AI Act** – the processing of special categories of personal data to detect and correct bias in high-risk systems (article 10, par. 5), and the further processing of personal data for developing certain systems in the public interest within a sandbox (article 59). Contributors stress that these provisions should not become a loophole: competent authorities should document the justification for such use, the type of data and domain involved, the model type, the alternative methods (including synthetic or pseudonymised data) that were tested and found insufficient, and the privacy safeguards applied.

2.4 Accountability, human oversight and agentic systems

Contributors repeatedly link accountability to transparency, explainability and effective human oversight, and warn that without a reliable record of who changed what, at which stage of the AI lifecycle, and with what understanding of the consequences, responsibility risks being assigned to the closest operator rather than to the actor best placed to prevent or mitigate harm. Sandboxes should therefore be **designed around explicit observation questions from the outset**: what level of understanding is expected of human operators at each lifecycle stage, what traceability and logging is required, and what technical limitations remain for different model types. This is considered especially important for complex and agentic systems capable of decomposing tasks or acting through an orchestrator, where providers should be required to test operational boundaries, define escalation thresholds and design feasible, timely intervention points. At the local level this concern is acute, since territorial authorities are increasingly deploying AI decision-support systems directly in citizen-facing services; sandboxes at this scale should test not only the technology itself but the real capacity of local public servants to understand, oversee and, where necessary, stop such systems.

2.5 Specialisation, ecosystem connectivity and institutional culture

Several contributions suggest that sandboxes could be usefully specialised not only by industry sector (e.g. automotive, healthcare) but also around **specific legal requirements** that cut across overlapping

instruments – for example human oversight, which appears in the AI Act, the GDPR, the Platform Work Directive and the Digital Services Act, or data governance obligations found across the AI Act and the GDPR. This kind of specialisation would allow regulators to develop deeper expertise and produce more fine-tuned guidance, which EU-level bodies could then consolidate.

To become "genuine" laboratories rather than a series of regulatory meetings, contributors call for **meaningful connectivity with the wider EU AI ecosystem** – TEFs, European Digital Innovation Hubs (EDIHs), AI Factories, and data spaces – including offers of technical testing and data access, not only compliance dialogue. The workshop discussion added that innovation is, ultimately, also a matter of institutional culture: for many authorities, sandboxing will be the first time they design and implement this kind of arrangement, which requires new resources and a shift in working practices as much as new procedures.

Specialisation can be territorial as well: a region with a wide range of expertise in, for example, healthcare, mobility or tourism can tailor a sandbox to the most useful AI use cases for its economy and public services. This aspect is also strengthened by the fact that much of the wider EU ecosystem – EDIHs, many TEFs and emerging local data spaces – is itself regionally anchored, making the regional level a natural point of connection between compliance dialogue and technical support.

3. Distributing responsibility across governance levels

Contributors broadly agree that the precise legal allocation of powers between authorities matters less to participants than having one clear, coordinated path through the sandbox – but they also agree that achieving this requires deliberate multi-level design, not improvisation.

3.1 A single point of contact, backed by a coordination network

From the participant's perspective, what matters most is a **strong coordination function** that brings in the relevant authorities only when genuinely required, with the depth of involvement of any given supervisory body proportionate to its relevance to the specific use case. Several contributions propose that each sandbox be supported by a national coordination network established at the design stage: a mapping of relevant authorities (market surveillance, data protection, sectoral regulators, fundamental rights bodies), designated contact points, fast communication channels and procedures for secure documentation sharing, with regular – for instance monthly – coordination meetings that do not require every authority's presence at every provider meeting. Under this model, the national competent authority (NCA) remains the primary point of contact and remains responsible for sandbox management as mandated by the AI Act, while gaining the means to coordinate efficiently with others.

Where a sandbox is established by a regional or local authority under Article 57, par. 2, this coordination network must also run vertically, connecting the territorial authority to the national competent authority, so that a locally-hosted sandbox does not become a parallel track disconnected from national conformity assessment and EU-level reporting. The mapping of relevant authorities should accordingly include the competent regional and local bodies alongside national competent authorities.

One contribution frames this as the need for a single **director of works**: one authority clearly in charge of coordinating input from data protection authorities, sectoral regulators and technical bodies, not primarily for efficiency but because it is the only realistic way to guarantee that participants receive coherent, non-contradictory guidance rather than navigating between authorities pulling in different directions. The

workshop discussion echoed this concern about fragmentation, with some participants advocating a single national AI Office to reduce the risk of regulatory arbitrage across different sandbox models.

3.2 Legal frictions and the limits of the sandbox's "non-supervisory" character

Contributors note that sandbox participation is generally non-supervisory in nature, but that this should not become a shield against addressing legal violations that surface during experimentation – for instance, where a provider enters a sandbox with a system trained on data that was collected unlawfully. In such cases, appropriate coordination with data protection authorities and lawful mitigation channels should still apply. This reflects a broader point made across contributions: sandboxes work best when they are understood as **spaces of coordinated regulatory responsibility**, in which each authority preserves its own mandate while participating in a shared learning process, rather than spaces where regulators temporarily suspend their ordinary roles.

3.3 Roles across levels of governance

A multi-level model recurs across contributions:

- **EU-level bodies (AI Office, EDPS)**: should establish common learning objectives and documentation standards, ensure consistency and comparability across Member States, disseminate good practice, and prevent divergent national interpretations – without micro-managing individual sandboxes. The AI Board's AI sandbox subgroup is repeatedly identified as the natural forum for this, alongside the Single Information Platform mandated by the AI Act (article 57, par. 17). The EDPS is seen as capable of playing an equivalent coordinating role for EU institutions and, through its role in the EDPB, of helping shape data-protection-sensitive sandbox methodologies; the EDPB, in turn, could play a role for GDPR aspects analogous to the AI Office's role for the AI Act.
- **National authorities**: should lead contextual supervision within their competence – market surveillance authorities leading on AI Act compliance, with data protection authorities structurally involved whenever personal data are processed, which will be the ordinary case for many high-risk systems, since data governance, profiling, human oversight, logging and explainability sit precisely at the intersection of the AI Act and the GDPR.
- **Regional and local authorities**: where they establish sandboxes under Article 57, par. 2 of the AI Act, they should be recognised as a distinct tier within this model rather than folded into the category of "national authorities". Operating closest to affected communities and to the public services in which AI is actually deployed, they are well placed to run contextualised, participatory experimentation responsive to location-based needs. Their role should be exercised within – and coordinated by – the national framework.
- **Sandbox participants**: should not be treated as passive recipients of guidance but as active producers of evidence, expected to document design choices, assumptions, organisational constraints, failures and mitigation measures, including why particular measures were adopted where the law leaves a choice (for example, the human oversight measures under Article 14 of the AI Act).

Contributors describe the objective, in one formulation, as **redistributing learning rather than redistributing authority**: accountability is strengthened less by re-drawing formal competences than by building shared responsibility for producing evidence and institutional knowledge. The workshop discussion added a concrete illustration of EU-level experimentation: a planned EDPS exercise to stress-test risk management,

transparency, fundamental rights impact assessments (FRIAs) and human oversight arrangements for EU institutions themselves.

3.4 Legitimacy, resourcing and legal accountability

One contribution situates this multi-level allocation of powers within a broader concept of **legitimacy**, understood as a precondition for accountability rather than a separate concern. From a vertical (EU/Member State) perspective, the EU already operates a multi-level governance model in which European bodies are responsible for communicating general objectives and the means of achieving them, gathering information about national authorities' activities, and feeding this back into national and subnational practice – an architecture that was confirmed through the AI Act's own legislative process and that, in this contributor's view, has in other domains produced a more mature and robust regulatory response than either a purely centralised or a purely national alternative.

From a horizontal perspective, the same contribution argues that data protection authorities should continue to play a central role and that AI sandboxes should **not be allowed to weaken personal data protection**, even inadvertently. It flags a genuine policy tension here: many national authorities already face constraints on human and financial resources, and there is a paradox in political demands for "less state" and "less administration" sitting alongside a regulatory architecture, sandboxes included, that cannot function without capable, well-resourced regulators – in other words, credible innovation policy presupposes capable regulation rather than substituting for it. It also argues that a **genuine accountability agenda** cannot rest on institutional learning alone and needs to provide legal remedies: individuals who are harmed by a sandboxed AI system without having been participants in the experiment should have a clear route to hold the State liable, and, given the "Digital Omnibus on AI" simplification proposals that extend the European Commission's own direct role in this area, an equivalent liability mechanism should apply at EU level too.

4. Evidence, learning, and the boundaries of regulatory relaxation

4.1 What sandboxes should be expected to produce

Contributors consistently argue that technical benchmarks, while important, are only one part of the evidentiary picture. A **layered approach to classifying evidence** emerges across contributions:

- A first layer distinguishes cases by type of entity (start-up, public body, research organisation, large company), type of dataset, type of model, intended purpose and domain of use, to avoid overly general conclusions.
- A second layer classifies the evidence generated into categories such as: ethical and fundamental rights evidence (risks to dignity, non-discrimination, autonomy, access to services, procedural fairness); socio-economic evidence (opportunities, barriers, market effects); legal-interpretation evidence (recurring compliance challenges where AI Act obligations intersect with the GDPR, sectoral law or fundamental rights law); socio-technical evidence (practical approaches to implementing legal obligations in specific contexts); and regulatory-learning evidence (identified rigidities, loopholes, and possible amendments or guidance).
- A third layer is the underlying evidence itself: metrics, acceptable trade-offs, documentation practices, human oversight arrangements and organisational routines, including for systems that fall outside the formal high-risk category but still raise significant ethical, legal, safety or rights concerns, such as emerging agentic AI behaviours.

One contribution cautions against the word "evidence" itself, arguing that it implies a standard of verifiable, reproducible scientific proof that a sandbox – lacking a research design in the first place – cannot straightforwardly deliver. On this reading, it may be more accurate to speak of **sandbox "outcomes" tailored to different audiences**, since sandboxes pursue several distinct aims at once (regulatory, participant, market and societal): technical benchmarks are primarily of interest to other technical actors; failure and success cases and citizen input are informative mainly for the market, including investors, and for society at large; and organisational procedures and rights-based narratives are primarily useful to society and to regulators. The same contribution adds that involving civil society in sandboxes is important, but that the EU should be consistent with its own ambition and guarantee affected groups the practical means to participate – financial support, accessible formats and adequate time to respond – rather than inviting participation on terms designed for well-resourced organisations.

Contributors give particular emphasis to organisational and human-oversight sandbox outcomes, since AI governance failures often do not depend on technical performance alone. Suggested indicators include error detection rates, override rates, justified deviations from algorithmic outputs, indicators of automation bias, training effectiveness, decision traceability, and the actual capacity of human overseers to challenge or stop a system; comparable indicators are proposed for data governance, minimisation and explainability. **Affected individuals and communities** are described as an important additional evidentiary source, particularly for systems affecting access to work, welfare, credit, education, migration, healthcare or public services, with inputs ideally gathered through the same data protection and fundamental rights impact assessment (DPIA/FRIA) processes that the sandbox may itself be testing.

4.2 What should not be used to justify regulatory relaxation

Contributors are consistent, and notably firm, on this point. Evidence that a requirement is difficult, costly or inconvenient to implement should not, by itself, be treated as grounds for lowering the requirement: difficulty more often signals a need for clearer guidance, better cross-authority coordination, shared infrastructure, or more realistic implementation pathways than for deregulation. Claims of economic benefit, efficiency gains, market competitiveness or innovation potential, taken alone, are similarly considered insufficient grounds for regulatory relaxation; **innovation remains an important objective** but should not override evidence on **rights protection, accountability and public trust**. The use of sensitive data should not be normalised merely because it may support bias mitigation – alternative methods, including synthetic data and privacy-preserving approaches, should be tested first and their limitations documented. One contribution adds a useful conceptual distinction: "simplification" (greater regulatory clarity, consistency and tailored guidance) is not the same as "regulatory relaxation" (lowering of substantive standards), and many sandbox initiatives have in practice led regulators to update and clarify guidance rather than to relax rules. Where innovation objectives genuinely conflict with rights protection or sectoral requirements, contributors suggest the issue should be assessed collectively with the relevant authorities and, where appropriate, escalated to the AI Office and other Member States rather than resolved unilaterally in favour of the innovator.

The workshop discussion reinforced several of these points in practical terms: participants noted that synthetic testing datasets were sometimes insufficient during actual experimentation, that regulatory learning needs to be understood more broadly than compliance learning, and that agile, benchmark- and threshold-based approaches to regulation were an area meriting further exploration – illustrated by a banking-sector exercise simulating climate-change scenarios that required cross-institutional data sharing.

4.3 The centrality of failure cases, participation incentives and the evidence base for sandboxes themselves

A further contribution adds a related but distinct proposal: rather than leaving the process open-ended, the legislator should fix a **clear protocol for how sandbox experiments are to be conducted** before they start, specifying both the process to be followed and the outputs expected. Within such a protocol, this contributor considers virtually every category of sandbox output legitimate to feed into policy discussion, with one specific exception – failed cases should not be used to argue for relaxation or simplification. This conclusion differs in its reasoning from, but arrives at broadly the same destination as, the wider consensus among contributors that difficulty or failure should prompt clearer guidance rather than lower standards. Indeed, multiple contributions single out **failure cases as especially valuable**, on the basis that a sandbox reporting only success stories is not performing its institutional function: unsuccessful experiments often reveal more about the robustness of governance arrangements than successful deployments, showing where safeguards are ineffective and which regulatory or organisational assumptions do not survive contact with real-world deployment.

Several contributions flag a further, more self-referential evidence gap: a risk that organisations decline to participate in sandboxes because the perceived costs of participation – time, resources, exposure – outweigh the perceived benefits, even where reputational or commercial advantages are available in principle. Contributors note that **there is currently little robust economic analysis quantifying the real benefits that participants obtain**, which one contribution frames as a genuine paradox given the EU's own preference for evidence-based policymaking: regulatory sandboxes have been extended across a wide range of legislative areas without much replicable proof of their effects. Competent authorities are encouraged to treat this gap in the evidence base for the instrument itself as seriously as the evidence the instrument is meant to generate about AI systems.

5. Access, transparency, participation and safeguards

Contributions to the fourth question – gathered primarily through the offline written round – converge on the view that governance design choices are not administrative detail but the substantive core of whether a sandbox supports trustworthy AI.

5.1 Access and selection criteria

Several contributions identify the selection of sandbox projects as, in itself, the **single most consequential design choice**: the acceptance framework determines whether a sandbox focuses on frontier cases that reveal future regulatory challenges, or on more mature applications with broader adoption potential, and this choice should be made deliberately rather than by default. "Societal relevance" recurs as a common access criterion but is criticised as often too subjective; contributors propose **more concrete factors**, including the potential impact – positive and negative – of the innovation on fundamental rights, the applicant's willingness to engage with affected stakeholders throughout experimentation, the degree of genuine legal, technical or organisational uncertainty raised by the use case, and public-interest relevance. In this connection, one contribution proposes a simple but concrete screening device: requiring project submissions to state explicitly which specific data-governance or accountability issue the project is designed to test, so that admission decisions are anchored in an identifiable regulatory question from the outset rather than in a generic account of the innovation.

Contributors caution against access criteria that **implicitly reward well-resourced actors** seeking a quasi-formal regulatory endorsement. A degree of organisational maturity – documented risk-management procedures, accountability structures, a clear articulation of intended purpose – is considered a legitimate

precondition for meaningful experimentation, but this should not translate into a barrier for smaller organisations. Several contributions call for active support to start-ups and SMEs so that lower resources do not lower accountability expectations. One contribution also flags the residual possibility, under the AI Act, for non-profit organisations to act as sandbox innovators where an AI system is deployed for a non-commercial but professional activity (for example, a charity or a trade union developing a system for its own community), suggesting this deserves more active consideration than it currently receives.

5.2 Transparency and exit reporting

Transparency is described as a feature that should run throughout the sandbox lifecycle rather than appear only at the end. Contributors converge on a **layered model**: confidential technical documentation for regulators, alongside structured, meaningful public summaries for wider dissemination. On exit reports specifically, several contributions argue that the relevant question is not whether to publish, but how: a public version should go beyond a mere summary and provide substantive information on how regulatory issues were addressed and how risks to health, safety and fundamental rights were identified and mitigated, since transparency in name only will not support genuine accountability or regulatory learning. Mid-term reporting is recognised as a powerful transparency tool that also carries real costs – interim findings are, by definition, partial and potentially sensitive, and managing what can be disclosed without compromising the participant or the process is a significant commitment, particularly for large or complex sandboxes.

5.3 Participation of affected communities

Contributors treat meaningful participation of affected communities as a governance mechanism in its own right, not an optional add-on, and stress that it should **begin at the design stage and continue throughout experimentation** – not be limited to collecting feedback once key decisions have already been made. Concrete practices suggested to include stakeholder mapping, consultations on both sandbox design and preliminary insights, advisory bodies, co-design methods drawn from the participatory and value-sensitive design literature, and collaborative governance arrangements that give rightsholder representatives (civil society organisations, trade unions, patient or worker representatives) a seat on the committees making relevant decisions. In sectors such as healthcare, the involvement of clinicians, caregivers and patient representatives is described as performing not only a consultative but a validation function, testing whether governance arrangements remain workable under real conditions of use. In this context, proximity gives regional and local authorities a structural advantage in meeting this expectation: affected citizens, local patient or worker representatives and place-based civil-society organisations are easier to identify, convene and genuinely involve at the territorial level than through national processes.

Contributors also emphasise that participation should include people who are not already knowledgeable about AI or regulation, since building trust requires **reaching beyond the usual expert audience**; workshop discussion added that a genuine mismatch remains between how citizens currently think, speak and participate in public processes and what civic engagement in a sandbox would actually require. As noted above, this participatory ambition also has a resourcing dimension: one contribution stresses that the EU should be consistent with its own aim of involving civil society by ensuring affected communities are actually equipped – financially and logistically – to take part, rather than being invited to participate on terms that assume the availability of resources only well-established organisations tend to have.

5.4 Safeguards for trade secrets and fundamental rights

Contributors are clear that **confidentiality cannot become a reason to prevent meaningful scrutiny** of systems that affect fundamental rights, while also recognising that legitimate trade secret protection

matters, including to sustain participants' incentive to take part. Proposed mechanisms include trusted intermediaries, structured multistakeholder advisory bodies, non-disclosure agreements and clear participation protocols, together with the layered public/non-public reporting model described above. One contribution notes, more optimistically, that carefully-communicated regulatory compliance can itself become a competitive advantage rather than purely a disclosure risk.

5.5 Documenting and disseminating outcomes

Contributors call for **standardised exit reports** that identify the legal questions tested, the evidence generated, the safeguards adopted, the failures encountered, the unresolved issues, and the guidance that can be generalised, feeding into AI Office and EDPS reporting, harmonised standards, common specifications, AI literacy initiatives, and future legislative or delegated-act reviews. Suggested formats include aggregated reports, anonymised case studies, thematic guidance, operational checklists and templates, a database cross-referencing sandbox projects to specific AI Act articles, and an annual EU-wide sandbox summit to build a genuinely shared cross-border learning culture, which several contributors consider currently absent. In a similar vein, one contribution recommends establishing a dedicated European repository of sandbox outcomes (in line with the single information platform) and asking the AI Board to produce periodic thematic analyses of the material collected, following the model used for many years by the Article 29 Working Party in the data protection field. For regional or local sandboxes, standardised reporting is precisely what allows locally-generated learning to become comparable across regions and to feed the EU-level repository; without it, regional experimentation risks producing valuable insights that never travel beyond the authority that generated them.

5.6 Measuring success

A final recurring theme is that evaluation criteria for sandboxes need more attention than they currently receive. Contributors argue that success should not be measured only by the number of systems that reached the market or demonstrated compliance, but also by the **quality of the institutional knowledge** generated, the **effectiveness of the safeguards** tested, and the quality of **stakeholder participation** achieved. Establishing evaluation criteria from the outset, rather than retrospectively, is seen as essential to strengthening accountability for the sandbox mechanism itself.

6. Beyond the sandbox: Towards a common standard for regulatory learning

A further suggestion looks beyond the sandbox instrument itself. On this view, sandboxes are only one of several parallel fora capable of generating regulatory learning about AI; others include stakeholder consultations on fundamental rights issues raised by specific applications, human trials, public procurement processes involving AI, self-certification schemes operated by public bodies, and shared compliance labs run through academic networks or AI Factories. The practical challenge is not a shortage of learning opportunities but the difficulty of efficiently discovering and exchanging the lessons that arise from these different, parallel episodes – quickly and consistently enough for the learning to matter.

The suggestion is to address this by developing a common governance template for regulatory learning of this broader kind, so that lessons generated in any of these fora – not only in sandboxes – can be found, compared and reused more quickly, thereby improving regulatory certainty and reducing the practical cost of AI Act compliance across Member States and sectors. Concretely, this could take the form of a management system standard for AI regulatory learning, built on the harmonised structure that ISO already uses for management system standards such as ISO/IEC 42001 (AI management systems) and the ISO/IEC 27000 family (information security management). Relying on an established, harmonised structure

would make such a standard easier for organisations to adopt alongside management systems they may already operate.

One practical route towards such a standard would be to use the CEN/CENELEC Workshop Agreement mechanism to build initial multi-stakeholder consensus on its content – a comparatively low-friction step that could, if it gained sufficient traction, feed into formal standardisation work through CEN/CENELEC's Joint Technical Committee on Artificial Intelligence (JTC 21).

7. Key recommendations for competent authorities

The following key recommendations (KRs) are addressed to national competent authorities and EU-level bodies that are establishing, or already operating, AI regulatory sandboxes under the AI Act. They are grouped thematically and are intended to be read together with the fuller discussion above.

Table 1. List of key recommendations for AI regulatory sandboxes

#	Description
	Institutional design and mandate
KR01	Give the sandbox an explicit mandate, and the corresponding resources, to generate transferable regulatory knowledge – not only to support individual participants toward compliance.
KR02	Treat the selection of projects admitted to the sandbox as a strategic governance choice: decide deliberately whether the sandbox will prioritise frontier use cases that surface future regulatory challenges, mature use cases with broader adoption potential, or a defined mix of both, and align this with national and EU-level priorities (for example, sectors of particular technological or societal relevance).
KR03	Invest in interdisciplinary operator capability – technical, legal, organisational and policy expertise combined – since even a well-designed governance framework will remain superficial without the capacity to engage deeply with complex use cases.
KR04	Set access criteria around genuine regulatory uncertainty, fundamental rights sensitivity and public-interest relevance rather than around applicants' resources; support smaller organisations (SMEs, start-ups, public bodies, research organisations, and, where relevant, non-profit innovators) in meeting reasonable organisational-maturity expectations rather than lowering those expectations.
KR05	Keep sandbox timelines predictable and limited; prolonged, open-ended engagements disproportionately disadvantage smaller organisations.
KR06	Build in connectivity with the wider EU AI ecosystem – Testing and Experimentation Facilities, European Digital Innovation Hubs, AI Factories, data spaces – so the sandbox can offer genuine technical testing and data access, not only regulatory dialogue, and communicate clearly how these initiatives interconnect.
KR07	Require project submissions to state explicitly the specific data-governance or accountability issue the project is designed to test and put in place a peer-review mechanism among the officials running the sandbox to help control for individual bias in assessment.
KR08	Identify comparable sandbox initiatives, assess what worked well, what did not, and what could be improved. Develop a benchmark based on lessons learned to inform the design

#	Description
	of the new sandbox. Regulatory experimentation should serve as a mechanism for generating and sharing knowledge across regulators, rather than creating isolated learning processes.
	Multi-level coordination
KR09	Establish a national coordination network at the design stage of the sandbox, mapping relevant authorities (market surveillance, data protection, sectoral regulators, fundamental rights bodies), designated contact points, fast communication channels, and procedures for secure documentation sharing.
KR10	Assess, early and case by case, which additional authorities need to be involved for a given application, and calibrate their involvement to the relevance of the use case rather than defaulting to full involvement of every authority in every meeting.
KR11	Maintain the national competent authority as the sandbox participant's single, consistent point of contact, while coordinating regularly (for example monthly) with other involved authorities behind the scenes.
KR12	Do not treat the non-supervisory character of sandbox participation as a reason to ignore legal violations that surface during experimentation; route them to the appropriate authority and lawful mitigation channel.
KR13	For Member States with complex or federal administrative structures in particular, consider a clearly designated coordinating authority (a "single point of coordination") responsible for ensuring participants receive coherent, non-contradictory guidance.
KR14	Ensure data protection authorities remain structurally involved rather than sidelined by resource pressure; resourcing constraints on supervisory authorities should be treated as a risk to the sandbox model itself and addressed accordingly, not absorbed silently.
KR15	Pair institutional learning with legal remedies: ensure that individuals harmed by a sandboxed system who were not themselves participants have a clear route to seek redress and extend equivalent liability arrangements to EU-level bodies where they take on a more direct operational role.
	Evidence generation and use
KR16	Define, in advance, the categories of evidence the sandbox expects to generate (technical, organisational, socio-economic, legal-interpretative, rights-impact), while leaving room to capture unanticipated governance issues as they emerge.
KR17	Actively seek and document failure cases and unsuccessful approaches; treat their absence from a sandbox's output as a signal that something is being under-reported, not that the sandbox is succeeding.
KR18	For human oversight and similar open-textured obligations, collect concrete operational indicators (e.g. override rates, error-detection rates, automation-bias indicators, decision traceability) rather than relying on self-reported compliance narratives alone.
KR19	Require documented justification whenever a sandbox relies on the AI Act's exceptional data-processing possibilities (e.g. Articles 10 and 59), including the alternatives tested and found insufficient and the safeguards applied.
KR20	Do not treat difficulty of compliance, cost, or claims of innovation and competitiveness benefits, on their own, as sufficient grounds for regulatory relaxation; use such findings instead to identify where clearer guidance, shared infrastructure or better coordination is needed.

#	Description
KR21	Distinguish explicitly between “simplification” (clarity, consistency, better-tailored guidance) and “regulatory relaxation” (lowering of substantive standards) when communicating sandbox outcomes, to avoid the two being conflated in public debate.
KR22	Fix, before experimentation begins, a clear protocol specifying how sandbox experiments will be conducted and what outputs are expected, so that later disputes about what counts as usable evidence are minimised; agree in advance that failed cases will not be invoked to justify relaxation or simplification.
KR23	Commission or support independent economic analysis of the actual costs and benefits of sandbox participation, to close the current evidence gap about whether the incentives for participants are adequate – applying the same evidence-based standard to the sandbox instrument that the instrument is meant to apply to AI systems.
Transparency and stakeholder participation	
KR24	Apply transparency obligations throughout the sandbox lifecycle, not only at exit; consider a layered model with confidential technical documentation for regulators and meaningful, substantive public summaries for wider audiences.
KR25	Ensure exit reports go beyond high-level summaries: they should explain how regulatory issues were addressed and how risks to fundamental rights, health and safety were identified and mitigated, while protecting legitimate trade secrets.
KR26	Build participation of affected communities into sandbox governance from the design stage, using concrete mechanisms such as stakeholder mapping, advisory bodies, co-design activities, and representation of rightsholders on decision-making committees, rather than treating participation as post-hoc consultation.
KR27	Deliberately seek input from people without prior expertise in AI or regulation, not only from organised civil society, to genuinely test whether affected communities understand and trust the process.
KR28	Use trusted intermediaries, non-disclosure agreements and structured advisory arrangements to reconcile trade secret protection with meaningful public accountability, rather than treating the two as inherently opposed.
KR29	Establish evaluation criteria for the sandbox itself before experimentation begins, covering not only compliance and market outcomes but the quality of institutional learning generated, and the quality of stakeholder participation achieved.
KR30	Back civil society and community participation with practical means – funding, accessible formats and adequate time – so that the EU’s own ambition to involve affected communities is not undermined by participation terms that only well-resourced organisations can meet.
EU-level learning and dissemination	
KR31	Report observations and open challenges regularly to the AI Board’s AI sandbox subgroup, rather than only at the end of a sandbox cohort, to support faster, harmonised interpretation of the AI Act across Member States
KR32	Make anonymised, practical, operational lessons learned publicly available via the AI Act’s Single Information Platform and other relevant channels, including synopses of subgroup discussions.

#	Description
KR33	Encourage the AI Office and EDPS to focus on ensuring comparability, consistency and re-use of sandbox outcomes across Member States, without micro-managing individual national sandboxes.
KR34	Consider a dedicated, recurring EU-wide sandbox forum (for example an annual summit) to share outcomes, surface common challenges, and build a cross-border learning culture that currently does not exist in a systematic form.
KR35	Where relevant, connect learning from AI Act sandboxes with adjacent experimentation – including data protection sandboxes run by data protection authorities and pilot exercises run by EU-level bodies such as the EDPS – to avoid duplicating effort or missing cross-cutting lessons.
KR36	Establish a European repository of sandbox outcomes and task the AI Board with producing periodic thematic analyses of it, drawing on the model the Article 29 Working Party used for many years in the data protection field.
Standardising regulatory learning across fora	
KR37	Recognise sandboxes as one instrument among several capable of generating AI regulatory learning – alongside stakeholder consultations, human trials, public procurement of AI, self-certification by public bodies, and shared compliance labs in academic networks or AI Factories – and avoid designing information-sharing mechanisms around sandboxes alone.
KR38	Explore developing a common management-system-standard template for AI regulatory learning, built on ISO's harmonised structure for management system standards (as used for ISO/IEC 42001 and the ISO/IEC 27000 family), so that lessons from different fora can be documented, discovered and compared consistently.
KR39	Consider using the CEN/CENELEC Workshop Agreement mechanism to build initial multi-stakeholder consensus on the content of such a standard, as a lower-friction step that could, if it gains traction, feed into formal standardisation work through CEN/CENELEC JTC 21.
Regional or local sandboxes	
KR40	Where regional or local authorities may establish sandboxes under Article 57, par. 2 of the AI Act, treat this territorial tier as a specific design choice: give it an explicit mandate, adequate legal and technical resourcing, and a clear articulation with the national competent authority.
KR41	Extend the national coordination network to include the competent regional and local authorities, and ensure that any territorially-hosted sandbox connects vertically to national conformity assessment and EU-level reporting, so that local experimentation does not develop as a disconnected parallel track.
KR42	Set minimum common conditions for regional and local sandboxes at national level – on admission, supervision, reporting and exit – while leaving room for place-based implementation, so that territorial flexibility does not fragment legal certainty
KR43	Use the proximity of regional and local sandboxes to strengthen the participation of affected communities, while backing local authorities and local civil society with the resources needed to make that participation meaningful.

#	Description
KR44	Ensure that outcomes from regional and local sandboxes are documented in a standardised, comparable form and fed into the European repository, and promote cross-regional exchange so that place-based learning becomes transferable.

8. Concluding remarks

The contributions gathered before, during and after the CPDP workshop point in a consistent direction: the regulatory value of AI sandboxes will be determined less by how many providers pass through them than by the quality, honesty and shareability of what is learned along the way. This requires competent authorities to treat sandbox design – access criteria, coordination arrangements, evidence categories, transparency duties, and mechanisms for community participation – as substantive policy choices in their own right, made deliberately and revisited over time, rather than as administrative formalities to be settled once and left unexamined. Done well, sandboxes can become genuine sites of institutional learning about how AI governance actually works in practice; done narrowly, they risk remaining a compliance service that leaves the broader regulatory and societal questions they were meant to help answer largely untouched.

Citation: CybeRights Center (2026), *Sandboxing Governance Models for Trustworthy AI and Data Use*, Policy note No. 1, https://www.cyberights.unifi.it/.