ERA Forum
https://doi.org/10.1007/s12027-026-00892-w

ARTICLE

# Personal data protection in experimental AI governance

**Aleš Završnik**[1] · **Saša Krajnc**[1]

Received: 31 March 2026 / Accepted: 14 July 2026


**Abstract**
The article examines the relationship between the European Union's Artificial Intelligence Act (AI Act) and the General Data Protection Regulation (GDPR), focusing in particular on the role of regulatory sandboxes as instruments for fostering innovation while safeguarding fundamental rights. The data-intensive nature of contemporary AI systems creates inherent tensions with core GDPR principles, such as purpose limitation, data minimisation, and the protection of data subject rights. Against this background, the article addresses three key questions: how the sandbox provisions of the AI Act should be reconciled with the GDPR; whether these provisions operate as *lex specialis* or rather as a contextual clarification of existing rules; and whether regulatory sandboxes can achieve a meaningful balance between innovation and data protection. The analysis shows that, if properly designed, sandboxes can support trustworthy AI through a cooperative and proportionate regulatory approach.

**Keywords** Personal data protection · Artificial intelligence · Regulatory sandbox · Frictions · EU AI act · GDPR

## 1 Introduction

The EU Artificial Intelligence Act (AI Act)[1] is often described as the world's first attempt at comprehensive, horizontal regulation of artificial intelligence (AI). It is also

[1] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13.6.2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013,

✉ A. Završnik
ales.zavrsnik@pf.uni-lj.si

S. Krajnc
sasa.krajnc@inst-krim.si

[1] Institute of Criminology at the Faculty of Law, University of Ljubljana, Poljanski nasip 2, SI-1000 Ljubljana, Slovenia

Published online: 03 August 2026

an experiment in how a legal order that already has a fundamental rights framework for personal data can accommodate the ever-expanding demand for data required for contemporary machine learning. The European Parliamentary Research Service (EPRS) study[2] on the impact of the General Data Protection Regulation (GDPR)[3] on AI, stresses that AI's need for data interacts in a particularly fraught way with fundamental principles of personal data protection, as established by the GDPR (like purpose limitation, data minimisation and the rights of data subjects). Scholars detected this friction at the advent of the contemporary AI revolution, Zarsky claimed GDPR is incompatible with big data methods.[4] However, EPRS also argues that the personal data protection principles can be interpreted in ways that do not preclude the development of socially beneficial AI, provided that the processing of personal data is constrained by appropriate safeguards. AI technologies risk large-scale surveillance, e.g. Lyon discusses how Big Data intensifies and expands contemporary forms of surveillance,[5] Murphy on new forms of "algorithmic Surveillance".[6] AI technologies also threaten fundamental rights (in various contexts),[7] which is why AI regulation should guarantee the protection of these rights when AI systems are developed and deployed. The AI Act's innovation chapter (Chapter VI) introduces specific provisions for *testing novel AI systems within regulatory sandboxes* (Articles 57 to 59), advancing the idea that if we want "trustworthy AI", regulators and developers need to learn together, *in situ*, rather than at arm's length and *ex post*.

This article aims to address three interlocking questions: *First*, how should the regulatory sandbox provisions of the AI Act be reconciled with the GDPR? Do they constitute *lex specialis* that modifies the rules of the GDPR, or are they better understood as a narrowly-tailored clarification of how those provisions apply in the context of public-interest AI development in sandboxes? *Second*, what are the friction points between the regulation of AI under the AI Act and the EU personal data protection legislation? And *third*, do regulatory sandboxes under the AI Act genuinely offer a suitable balance between innovation and data protection and what safeguards, like data protection impact assessments and continuous adaptation of DPIA, are needed to ensure that AI sandboxes do not become experimental zones of diminished accountability for participants with detrimental effects for data subjects' rights on the one hand and an administrative box-ticking exercise without any real value for AI developers and deployers on the other hand?

(EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act).

[2] European Parliamentary Research Service (EPRS) [9].

[3] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27.4.2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data and repealing Directive 95/46/EC (General Data Protection Regulation).

[4] *Zarsky* [19].

[5] *Lyon* [13].

[6] *Murphy* [16].

[7] For instance, in the context of criminal justice: *Završnik* [20]. In the workplace context: *Sprague* [18].

## 2 AI regulatory sandboxes under the EU AI act

The notion of a regulatory sandbox predates AI by almost a decade and was born in the financial sector. AI regulatory sandboxes can be seen as controlled spaces in which authorities and AI system developers can experiment with high-risk AI systems, while temporarily deviating from ordinary legal regimes and, in the process, generating valuable information for future regulatory adaptation. Sandboxes are a way to overcome the traditional shortcomings of the regulation of new technologies, especially where *ex ante* rules risk being either outdated or over-broad by the time they are applied.[8] While many different types of regulatory sandboxes exist, some common characteristics are typically present, namely: their temporary nature, the involvement of both regulators and companies, the waiving of existing legal provisions (and related liability) and the provision of tailored legal support for a specific project from the regulator.[9]

Article 3(55) of the AI Act defines a *regulatory sandbox* as a controlled framework set up by a competent authority which offers providers or prospective providers of AI systems the possibility to develop, train, validate and test, where appropriate in real-world conditions, an innovative AI system, pursuant to a sandbox plan for a limited time under regulatory supervision. This codifies several features already implemented in the digital health data and FinTech domains sandboxes:[10] the involvement of a public authority as convenor and supervisor; the focus on pre-market development and validation; the possibility of testing in real-world conditions; and the strict temporal limitation and ex ante agreement of a sandbox plan describing objectives, safeguards and exit conditions.

Article 57 of the AI Act requires every Member State to ensure that its competent authorities establish *at least one* AI regulatory sandbox at national level, operational by 2 August 2026. Sandboxes may also be set up jointly by several Member States and, for Union institutions and bodies, by the European Data Protection Supervisor. The AI Act thus transforms what had been a scattered, experimental technique into a mandatory component of the Union's AI governance infrastructure in which providers may *develop, train and test* innovative AI systems, potentially also using real-world data and operating in real-world conditions, under the supervision of competent authorities and on the basis of a sandbox plan (as defined in Art. 3(55) and Art. 57 of the AI Act).

To avoid fragmentation between national legislations, Article 58 mandates that the Commission adopt *implementing acts* to specify common principles on eligibility and selection criteria for participation in the AI regulatory sandbox, procedures for application, participation, monitoring and exit, as well as the terms and conditions applicable to participants. Those implementing acts must, *inter alia*, ensure that sandboxes are open, with transparent and fair criteria, to any provider meeting the conditions; procedures are simple, easily intelligible, and clearly communicated, and

[8] *Baldini and Francis* [2], p. 3.

[9] OECD [17].

[10] See Commission nationale de l'informatique et des libertés (CNIL) and Information Commissioner's Office – ICO [5].

access is free of charge for SMEs and start-ups except in case of exceptional cost recovery; sandbox participation and its legal effects are mutually recognised across the Union; and that sandboxes facilitate access to tools and infrastructure for testing, benchmarking and explaining relevant dimensions of AI systems (accuracy, robustness, cybersecurity, risk-mitigation).

## 3 Personal data processing in the AI regulatory sandbox

AI systems depend on large, diverse and continuously updated data sets for training, validation and deployment. AI development is inherently data-driven, as algorithms cannot accurately learn from their environment without large amounts of data, which often includes personal data as a fuel and/or the (by)product of AI applications.[11] Simultaneously, the personal data protection regulation subjects processing of personal data to strict regulation and established principles, like lawfulness, purpose limitation, data minimisation, storage limitation, fairness, transparency and accountability.[12] *Baldini and Francis* rightly note that, despite the absence of explicit sandbox regulation, the GDPR will apply to most AI sandboxes because high-risk AI almost inevitably leverages personal data during its development and operation.[13] Many AI applications process personal data both to build models and to make inferences concerning particular individuals, so that all kinds of personal data can be used to analyse, forecast and influence human behaviour, transforming both the data and their outputs into valuable commodities.[14]

Conversely, AI sandboxes rarely involve pure technical experimentation and possibly entail intensive processing of personal data – an issue that has been identified by the EU legislator. Article 2(7) of the AI Act recognises that the AI Act shall not affect the applicability of EU data protection law, including the GDPR. For the processing of personal data in the sandbox, Article 59 of the AI Act introduces a distinctive *lex specialis* legal regime: personal data lawfully collected for other purposes may be further processed within an AI regulatory sandbox solely for the development, training and testing of AI systems, under specific conditions. The AI Act adds a new exception to the 'purpose limitation' principle. While the AI research could be interpreted as a part of the existing 'research exception' from Article 5(1)b of GDPR, according to which, further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes according to Article 89(1) is not to be considered as incompatible with the "purpose limitation" principle, the AI Act introduced a new *lex specialis* provision in relation to AI R&D.

If a provider of an AI system wishes to benefit from this *lex specialis* regime, the AI system must safeguard *substantial public interest* in a limited set of fields, such as public safety and health, environmental protection, energy sustainability, transport safety, the protection of critical infrastructure, and certain public-administration

[11] *Mitrou* [14], p. 7.

[12] *Jedrzejczak* [12], p. 91.

[13] *Baldini and Francis* [2], p. 2.

[14] EPRS [9], p. I.

services. While, in principle, the public interest will in most cases be easily demonstrated, it remains to be seen how the public interest will be interpreted in practice (see the informative opinion on the public interest as a ground for the "research exception" to the purpose limitation principle in the EDPB study).[15]

Apart from pursuing the public interest, controllers must also be able to demonstrate their objectives can not be achieved with solely anonymised or synthetic data. Moreover, processing must take place in a secure environment (as stipulated in Art. 59(1)(d)); data may not be used to take decisions or other measures with respect to the data subjects; and the data must be erased at the end of the sandbox process. These conditions are very strict, and some scholars[16] hence plausibly doubt that availing of the exemption in Article 59 could be too burdensome to attract companies to engage in the AI sandbox regime.

One can also argue that the *catalogue of public-interest domains* is too narrow, as it does not explicitly include, for example, sandboxes aimed at socially beneficial AI in education, deliberation processes, housing, labour markets or cultural services, even though these are sectors where AI-driven profiling and decision-making can profoundly affect equality and autonomy of data subjects. An excessively narrow list of domains for sandbox-eligible AI systems may push developers to pursue the same experimental uses of personal data outside the sandbox framework, where AI Act's safeguards do not apply (Art. 59), which can lead to more severe infringements of data subjects' rights from the outset.

Given the new exception to the GDPR purpose limitation principle, as discussed above, one can also argue that Article 59 establishes a completely novel personal data processing regime, which applies only to sandbox use. However, the AI Act explicitly states in its recitals (140) and provisions (Article 59(3)) that it is without prejudice to EU and national data protection law, which remains fully applicable to AI systems that process personal data. This reference to existing data protection laws suggests that Article 59 is intended to give a more specific answer to some of the questions that the GDPR leaves open regarding further processing and compatibility, especially in data-intensive, experimental AI contexts.

The question, therefore, is whether Article 59 should be read as a genuine *lex specialis* vis-à-vis the provisions of the GDPR, or merely as a codified example in a high-stakes, experimental context. Article 59 of the AI Act presupposes that the original collection of data complies with the GDPR and that the subsequent sandbox use is merely a further processing of already collected (processed) personal data. Controllers cannot 'launder data' that was unlawfully collected or is unlawfully stored or excessive by bringing them into the sandbox under the guise of innovation. Nor does Article 59 supply its own legal basis, as it entails no provisions about consent, contract, legitimate interests, or public-interest/public-authority grounds. Any controller seeking to rely on Article 59 must therefore first show that the original dataset complies with the established legislative regime for personal data processing, and that further processing in the sandbox can be anchored in one of the legal bases under Article 6 GDPR. This implies that Article 59 offers merely a specific compatibility

[15] European Data Protection Board. (2021) [8].

[16] *Ahern* [1], p. 18.

framework for a narrow class of public-interest AI experiments, but does not displace the need for a GDPR-compliant legal basis and is not a legal basis on its own.

## 4 Friction points: the risks to personal data protection in the sandbox

### 4.1 Purpose limitation, data minimisation and storage limitation

According to GDPR's fundamental principles of purpose specification and limitation, data minimisation and storage limitation (Art 5(1)(b), (c) and (e) of the GDPR), e.g. personal data should only be processed for defined and clear purposes and only for these purposes it had been collected unless another lawful basis is clearly established (e.g. consent, legitimate interest, etc.). The aim of these principles is to minimise the collection and processing of data and the amount of information a data controller holds about an individual to ensure that the way the information is processed is consistent with the expectations data subjects may have in a democratic society.

The AI development pipeline destabilises and threatens each of these principles.[17] The point of large-scale machine learning is to extract meaning from data beyond what it was initially collected for, which presents a significant challenge to the purpose limitation principle. In practice, controllers often engage or, rather, 'slip' into excessive data collection, beyond what is immediately necessary, using overly broad definitions in personal data protection documents that secure technical compliance while undermining the underlying rationale of collection limitation.[18]

The data minimisation principle adds another layer of complexity. Formally, only personal data, which is adequate, relevant and limited to what is necessary for the specific purpose, should be processed. Yet the approach AI developers often pursue does not account for the constraints imposed by this principle, especially when the system's performance improves exponentially with more training data. In turn, such 'data hunger' of AI systems may lead into argument that any additional data contributes further to a more robust AI tool and is hence compliant with the purpose limitation principle, which leads to a vicious circular logic – more data being processed may still be compliant with purpose limitation principle as purpose has not changed, and also with the data minimisation principle, as more data in fact is technically needed to achieve the goal in the first place. More specifically, where to set the bar of the minimally needed language corpora for the development of semantic and language technologies, can only be relative and defined in relationship of one language corpora to the other, e.g. Slovenian language is considered a low-resource language (LRL), one that lacks sufficient digital data, linguistic tools, and online presence for developing robust AI like speech recognition or machine translation, but only compared to English, Spanish, French and German, for which larger datasets are available and better linguistics tools exists.

EPRS (The European Parliamentary Research Service) treats purpose limitation and data minimisation as the core conceptual friction points between GDPR and the

[17] *Zarsky* [19].

[18] *Chałubińska-Jentkiewicz and Nowikowska* [4], p. 188.

AI Act. They observe that the admissibility of processing personal data for new and different purposes has become crucial in the era of AI and big data, where repurposing and merging data from different sources and processing such data to address issues that were not contemplated at the time of collection are key.[19] The Service's interpretation of Article 6(4) of the GDPR seems to suggest that repurposing is permissible if two conditions are met: (1) the new processing must be compatible with the purpose for which the data was collected, and (2) processing must have a legal basis (which may be different from the original one). Reuse for a merely statistical purpose is generally compatible, provided adequate safeguards are in place and it does not affect the data subject individually.[20] The purpose limitation principle is not a goal in itself. GDPR[21] allows exceptions from data subjects' rights for statistical purposes, but also for research and archival purposes (Article 5(1)b) in combination with Article 89 of GDPR). Several recitals in GDPR directly acknowledge that data subjects' rights are not absolute rights and competing societal interests exist, e.g. "by coupling information from registries, researchers can obtain new knowledge of great value with regard to widespread medical conditions such as cardiovascular disease, cancer and depression." (Recital 157 GDPR). "To meet the specificities of processing personal data for scientific research purposes, specific conditions should apply in particular as regards the publication or otherwise disclosure of personal data in the context of scientific research purposes." (Recital 159 GDPR). While there is no explicit mention of artificial intelligence research, it can be assumed that AI research in the medical or any other scientific area is covered and can clearly advance the well-being of individuals and societies at large.[22]

With regard to storage limitation, Article 5(1)(e) of the GDPR prohibits keeping personal data when it is no longer needed for the initial purposes of the processing. There is a clear tension between the need for processing of large sets of personal data for AI development purposes on the one hand, and the principle of storage limitation on the other hand, as AI developers are tempted to keep training and operational data indefinitely in case future models or re-training cycles might require it, but also for the purposes of ensuring transparency and reproducibility. The data is still considered valuable as 'gold', as typically only model weights are disclosed, not the training data themselves – the process being described as 'open-washing' risk in the current AI landscape, where models release weights while hiding training data, claiming 'open source' status without meaningful transparency.

[19] EPRS [9], p. 52.

[20] EPRS [9], p. 52.

[21] Similarly, the Data Act mentions the possibility of transferring personal data to research organisations. See Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13.12.2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act).

[22] From Recital 157 of the GDPR: »..research on the basis of registries enables researchers to obtain essential knowledge about the long-term correlation of a number of social conditions such as unemployment and education with other life conditions. Research results obtained through registries provide solid, high-quality knowledge which can provide the basis for the formulation and implementation of knowledge-based policy, improve the quality of life for a number of people and improve the efficiency of social services. In order to facilitate scientific research, personal data can be processed for scientific research purposes, subject to appropriate conditions and safeguards set out in Union or Member State law.«.

### 4.2 Purpose creep and re-identification

AI tools and functionalities are likely to cause the processing of personal data for purposes other than those for which the data was originally gathered, but also for purposes that exceed those for which the system was originally programmed.

Processing personal data for purposes other than those for which it was initially collected is more the rule than the exception in the development of AI tools. The core idea of AI tools is to disrupt existing domains and workflows in the name of effectiveness and efficiency: 'doing more with less' is the norm. The AI tools are intended to disrupt data flows and tear down the walls of data silos, which were built for distinct purposes. Excavating new meaning from data is the very core idea and the selling pitch of AI machinery – and this can only be done if data is aggregated and combined. The meaning of data may not be insightful on its own; its value increases exponentially when aggregated and combined with other datasets at scale. The underlying assumption of the "'data is the new oil' metaphor is precisely that data, if stockpiled in large amounts, can garner new insights into any X (X may be disease, global warming, weather, stock trading, personal sympathies, etc). Disruption of AI is then simultaneously a disruption in the understanding of data's meaning. Excavating new meaning from data is precisely using the data collected for one purpose to garner insights for other purposes. Purpose and function creep is then a pressing concern whenever data and data silos – which were kept separate to protect individuals and their interests – are used to deliver the 'new insights'.

Through the use of AI algorithms, the identifiability of data subjects has vastly increased[23] as AI systems establish new links not only to obvious identifiers but also to metadata and background data (such as location, search history, preferences), as well as to biometric data such as voice or facial patterns. The personal nature of a data item becomes a contextual feature: an apparently anonymous piece of information may become personal data when an AI system combines it with additional personal data that enables re-identification.[24] It is no wonder that the definition of personal data can only be a open and relative as with sufficient computing power, any data may become personal data at some point in the future. According to GDPR, personal data is defined as "any information concerning an identified or identifiable natural person" (Preamble (26)), and to determine whether a natural person is "identifiable", account should be taken of all the "means reasonably likely to be used." Such an open definition may be able to accommodate an elusive definition of personal data, but it makes clear that it is reasonable to expect that the existing personal data protection regime may, along with the increasing power of AI, absorb all sorts of data in the future. There is a risk of re-identification as AI capabilities advance, potentially turning seemingly impersonal data into personal data. However, such a definition may progressively absorb all impersonal data and, in the long run, completely paralyse the effective protection of personal data – when all data becomes personal data, compliance and enforcement become impossible (or at least completely arbitrary).

Famously, e.g., the authors of anonymised movie ratings on Netflix could be re-identified by linking them to non-anonymous ratings in Amazon and IMDb

[23] EPRS [9], p.37.

[24] EPRS [9], p. 37.

databases.[25] The EU Commission's White Paper on AI hence rightly points out that AI may retrace and identify the data subjects from the anonymised data, thus processing data that were originally processed under another legal framework or even datasets that per se do not include personal data.[26] This is not a hypothetical risk but a structural possibility whenever powerful models are trained on rich data sets and then applied across contexts.

### 4.3 Automated decision-making

Article 22(1) of the GDPR prohibits any completely automated decisions based on profiling and having legal or significant effects on the data subject. This prohibition is mirrored by Article 59(1)(f) and Recital 71 of the AI Act, stating that the processing of personal data in the sandbox should not lead to automated decisions affecting data subjects. Still, Article 22(1) GDPR leaves ample room for automated decision-making. Firstly, the general prohibition does not apply when the AI system is used as a decision-support tool for human decision-makers, who autonomously decide whether to accept or reject the system's suggestions. In other words, the decision must not be based "solely" on automated processing. Additionally, Article 22(2) of the GDPR provides three exceptions to the general prohibition (when necessary for a contract, with legal authorisation, or with explicit consent).

Many decisions made today by AI systems fall within the scope of Article 21(1) of the GDPR, as AI algorithms are increasingly deployed across the public and private sectors. The ECJ has also been progressive in its interpretation of data subjects' rights. In the *SCHUFA Holding* Case C-634/21, the outsourcing of credit scores still does not mean that the lender relying on such a credit score was not carrying out automated decision-making as the Court stated that significant preliminary decisions also fall within the scope of Art. 22 of the GDPR, provided that the main decision "heavily relies" on them.[27] SCHUFA's credit score played a "determining role" in the granting of credit and therefore the "establishment of that value" in itself constituted a "decision" within the meaning of Article 22(1) GDPR.

The use of AI makes it more likely that a decision will be based "solely" on automated processing, as humans may not have access to all the information that is used by AI systems, and may not have the ability to analyse and review the way in which this information is used.[28] Thus, rather than being able to demand that no automated decision be made about him or her, the data subject will more likely only be entitled to an explanation as to how the automated decision has been taken. In the *Dun & Bradstreet Austria* Case C-203/22, the ECJ confirmed the right to explanation of automated decision-making under the GDPR. The ECJ court clarified the expressions "meaningful information" and "logic involved". Explanation must be "meaningful", which always means – or implies – that the information provided under Article 15(1)(h) must be, at the same time, functional, important, relevant, and

[25] See https://courses.csail.mit.edu/6.857/2018/project/Archie-Gershon-Katchoff-Zeng-Netflix.pdf.

[26] European Commission [7], p. 11.

[27] *Müller* [15].

[28] EPRS [9], p. 60.

intelligible. The explanation of the "logic involved" referred to in Article 15(1)(h) covers all relevant information concerning the procedure and the principles of a specific result. The Court established the principle of algorithmic transparency in the GDPR, and, for the first time, it expressly linked the right of access to the rights to react to automated decisions, enshrined in Art. 22(3) GDPR.[29]

For data controllers, such transparency obligations might not be completely straightforward. *Mitrou* notes that some of the most powerful AI tools are also the most opaque, and that the complexity of AI-based processing makes it particularly difficult to ensure transparency.[30] *Chałubińska-Jentkiewicz and Nowikowska* warn that processing within the AI system may result in a complete loss of control of personal information, and that it may prove difficult to carry out controls and audits because the complexity of some algorithms will require research by sufficiently competent and trained individuals.[31]

The EPRS study proposes a set of *ex ante and ex post* obligations for data controllers in cases of automated decision-making: providing information on the existence of profiling, i.e. on the fact that the data subject will be profiled or is already being profiled; general information on the purposes of the profiling and decision-making; general information on the kind of approach and technology that is adopted; general information on inputs (predictors) and outcomes (predictions); general information on the importance of input factors in determining the outcomes; specific information on what data have been collected about the data subject and used for profiling; specific information on what values for the features of the data subject determined the outcome; and what specific information had been inferred about the data subject.[32]

The nature of personal data processing within the AI system might be aligned with the GDPR's conception of statistical processing, where the result of processing is not personal data, but aggregate data, and that this processing is not used in support of measures or decisions regarding any particular natural person.[33] In principle, sandboxed development and testing should resemble such statistical processing: real personal data are used, but only to produce general models and metrics, not to target individuals.

### 4.4 Ethical and social risks

AI entails many potential risks beyond personal data infringements, such as racial disparities, gender-based and other forms of discrimination, misuse for criminal purposes,[34] and misleading children into committing self-harm.

Sandbox experiments, especially those that draw on health, welfare, or policing databases, may disproportionately expose certain groups or communities to AI

[29] More in https://www.europeanlawblog.eu/pub/lwchuopd/release/1.

[30] *Mitrou* [14], pp. 7, 39.

[31] *Chałubińska-Jentkiewicz and Nowikowska* [4], p. 188.

[32] EPRS [9], p. 64.

[33] Recital 162 GDPR.

[34] Europol [10].

scrutiny. In predictive policing literature, risks of under-policing and over-policing have been documented:[35] minority groups are being policed more than others, which leads to more crime being discovered and more data being produced, feeding the algorithms to increasingly flag these communities and their members as riskier. On the flip side, such a "vicious circle effect"[36] means that allocating police resources to minority communities leads to fewer resources and under-policing of other communities and other types of crimes. In the social welfare domain, UN Special Rapporteur Alston showed how the digitisation and automation of social welfare benefits lead to a "digital welfare dystopia," in which recipients are increasingly under surveillance and harassed.

These risks cannot be mitigated merely through personal data protection laws and are sometimes invisible at the individual level.

However, a potential benefit of the regulatory sandboxes is that testing can reveal and highlight biases in datasets and models, which can then be mitigated. AI may also be used in ways that exploit less informed or less vigilant users. It is already used for phishing and other illicit online behaviour, when online scammers mimic human conversational traits or even completely impersonate users, e.g. with AI cloning of voice, and can use AI to great effect to mislead users[37] and steal personal data from unassuming individuals.[38] In such cases, digital sandboxes can be used to inform policymakers and the public and increase digital literacy.

AI systems may also unintentionally produce harmful consequences despite the best efforts. In the crime control domain, scholars have clearly shown how including "crime history" of an individual to calculate their risk score of re-committing a crime leads to racial disparities (Harcourt) despite data on race being explicitly excluded from algorithmic processing. The racial disparity is not an intended consequence, but still, the AI will produce racially disparate effects. Similarly, algorithms may collude to set prices and violate antitrust laws; see the algorithmic collusion case in the USA, where the DOJ and FTC are identifying harms to consumers (e.g., in the Hotel Room Algorithmic Price-Fixing Case in 2024).[39]

Addressing inherent social risks should be a task for both the private and the public sectors. The Maastricht University UM-DPCSR Framework, comprising five principles, 25 rules, and 44 controls on ethical and socially responsible data processing, insists that companies and organisations must start to consider data protection and cybersecurity as assets and moral imperatives rather than mere legal compliance obligations.[40] On the other hand, government institutions already use AI in many areas, and stand to benefit from the further adoption of these technologies. Law and policymakers have a significant role in shaping how AI technologies impact citizens'

[35] *Egbert; Leese* [6].

[36] *Završnik* [21].

[37] Europol [8].

[38] *Chałubińska-Jentkiewicz and Nowikowska* [4], p. 184.

[39] US Hotels Accused of AI-Driven Price Collusion via Algorithmic Platform: https://www.ftc.gov/news-events/news/press-releases/2024/03/ftc-doj-file-statement-interest-hotel-room-algorithmic-price-fixing-case.

[40] See https://www.maastrichtuniversity.nl/research/maastricht-european-centre-privacy-and-cybersecurity/corporate-social-responsibility-csr-4.

lives through regulation, policy, and best practices.[41] The EU High-Level Expert Group on AI (AI HLEG) envisaged that AI regulatory sandboxes could build capacity around fundamental rights impact assessments for experimental AI applications,[42] as AI tools can, in principle, lead to the violation of any fundamental right.

### 4.5 Oversight conflicts

According to the provision of Article 57(10) of the AI Act, in cases where personal data is processed, both data protection authorities (DPAs) and AI Act surveillance authorities can claim jurisdiction over sandbox participants. In several EU Member States, DPAs are already acting as AI regulators, often in cooperation with other regulatory bodies (for example, telecommunications regulators).[43] The EU Commission's White Paper[44] similarly insists that AI requirements must be effectively enforced by competent national and EU authorities, who should be in a position to investigate individual cases, and also to assess the impact on society. Such dual authority can be both a setback and an opportunity in the regulatory sandbox. Participants can benefit from coordinated guidance from both the AI Act and GDPR competent authorities, and obtaining direct advice from enforcement authorities can give participants a compliance advantage over competitors.[45] At the same time, this situation creates regulatory uncertainty, presenting providers of AI systems with possibly conflicting requirements issued by multiple authorities. For sandbox learning to scale, competent authorities should exchange best practices and align their approach toward sandbox participants.

## 5 Designing sandboxes with personal data protection safeguards

### 5.1 DPIAs and data protection by design and by default

AI systems might be trained on real personal data from the very beginning of their development, and it is safe to assume that many AI systems which apply for sandbox testing have been using data from real individuals or might indeed have already been deployed in real-world conditions. Given the characteristics of machine-learning systems and the aim of regulatory sandboxes, it is highly likely that most AI tools tested in a regulatory sandbox setting will also fall into the high-risk category as defined by the AI Act. The GDPR's risk-mitigation instruments like Data Protection Impact Assessments (DPIAs), prior consultation with DPAs, and continuous monitoring are concrete mechanisms that should be considered when developing novel AI tools, especially in cases of high risks to data subjects.

[41] *Chałubińska-Jentkiewicz and Nowikowska* [4], p. 190.

[42] EU High-Level Expert Group on AI, Policy and investment recommendations for trustworthy Artificial Intelligence, 2019, para. 29.2.

[43] *Baldini and Francis* [2], pp. 4-5.

[44] European Commission, White Paper on Artificial Intelligence [7], p. 23.

[45] *Baldini and Francis* [2], p.6.

The GDPR requires a DPIA when the processing of personal data is likely to result in a high risk to the rights and freedoms of natural persons, especially where processing involves a systematic and extensive evaluation of personal aspects based on automated processing, including profiling (Article 35). The DPIA must be drafted prior to undertaking such processing and must be updated when new features or modifications are introduced, or new purposes are pursued (Article 35(11)). DPIAs thereby become part of a broader risk-based approach to data protection, intended to redirect regulation toward *ex-ante* risk management practices.

In the context of AI regulatory sandboxes, the question arises about the role of the DPIA in sandbox participation when personal data is processed. Here, several options are at hand: (i) a DPIA could be a mandatory requirement for sandbox participation, (ii) the participants who had carried out a DPIA prior to applying for the sandbox would be more eligible for selection (would be awarded higher scores for the inclusion in the sandbox) or (iii) a DPIA would need to be conducted during sandbox testing, if the regulatory authority deems it necessary, with the penalty of suspension or termination for non-compliance.

Certain stakeholders, like the Centre for Information Policy Leadership (CIPL), whose members are predominantly large multinational companies, suggests that a DPIA should be seen primarily as a means of assessing risk, and should not be a precondition for participation in the sandbox, but sandbox testing with pilot data, under supervision, can usefully anticipate or integrate a DPIA and may justify repeating the DPIA after sandbox participation.[46] On the other hand, *Baldini and Francis* highlight that ethical principles and safeguards going beyond the explicit legal requirements are fundamental for preventing harm, because sandboxes may allow deviations from existing law, and harms may be inflicted upon the fundamental rights and freedoms of individuals.[47] This position could speak in favour of a more stringent approach, where a DPIA would be a prerequisite for sandbox participation.

Regulatory experimentation is not an alternative to risk assessment, but a context in which DPIAs, privacy-by-design, and accountability should be carefully considered, because the technology and its risks are, by definition, unclear at that point in time and less well understood. The EPRS study[48] concludes that the GDPR's provisions on privacy by design and default do not hinder the development of AI systems, if correctly designed and implemented, though they may entail some additional costs. It needs to be clarified which AI applications present significant risks to personal data protection and therefore require a preventive DPIA, perhaps beyond Art. 35, and possibly the preventive involvement of data protection authorities.[49] DPIAs and data protection by design only work in an ecosystem of supervision.[50] Such position is reflected in Article 57(10) of the AI Act, which stipulates that DPAs must be involved in any AI regulatory sandbox project that in volves the processing of personal data.

[46] Centre for Information Policy Leadership (CIPL) [3], p. 15.

[47] *Baldini and Francis* [2], p. 8.

[48] EPRS [9], p.2.

[49] EPRS [9], p. II.

[50] EPRS [9], p. 68.

### 5.2 Safeguards within the AI act

The AI Act contains further safeguards for the processing of personal data in the regulatory sandbox (Article 59(b) to (j)):

1.) Necessity and data substitutability: real personal data are allowed only where less intrusive data types (anonymised, synthetic or other non-personal data) are insufficient;
2.) Functional separation and sharing limitations: Any personal data processed in the sandbox must be kept in a functionally separate, isolated and protected data processing environment under the provider's control, with access restricted to authorised persons. Any personal data created in the sandbox cannot be shared outside the sandbox;
3.) No-decision rule: processing of personal data in the context of the sandbox may not lead to measures or decisions affecting the data subjects or the application of their rights on the protection of personal data;
4.) Deletion: personal data processed in the context of the sandbox must be protected with appropriate technical and organisational measures and deleted once the participation in the sandbox has terminated or the personal data has reached the end of its retention period;
5.) Transparency: a complete and detailed description of the process and rationale behind the training, testing and validation of the AI system must be kept together with the testing results as part of the technical documentation.

### 5.3 Liability exemption

The AI Act allows for a certain amount of regulatory flexibility within sandbox testing. Provided that sandbox participants observe the specific plan and the terms and conditions for their participation in a specific sandbox and follow in good faith the guidance of competent authorities, no administrative fines should be imposed by the authorities for infringements of the provisions of the AI Act (Article 57(12)).

The AI Act also imposes obligations on competent authorities. Enrollment in regulatory sandboxes requires authorities to provide guidance, supervision and support to participants, including guidance on regulatory expectations and how to fulfill the requirements and obligations of the AI Act. Where competent authorities responsible for specific sectoral legislation (for example, telecommunications, banking, insurance, healthcare, personal data protection, etc.) were actively involved in supervising the AI system in the sandbox and provided guidance on compliance, no administrative fines should be imposed under that specific legislation. This provision (Article 57(12)) implies, that if competent market surveillance authorities issued guidelines and were actively cooperating with the sandbox participant, not only can a loyal participant not be fined for infringements of the provisions of the AI Act, but can also not be fined for the infringement of the sectoral legislation.

Confidentiality concerns, the inability to relax legal rules during the sandbox, and the inability of the sandbox to deliver a presumption of conformity with the AI Act

are identified as pertinent concerns for innovators.[51] The liability exemption thus addresses an important, if not essential, aspect of sandbox design. For a sandbox to be attractive to participants, it must offer clear benefits that outweigh potential risks and downsides, such as the disclosure of trade secrets and exposure of participant AI's flaws and unmitigated risks. The prohibition on administrative fining creates a 'safe space' that allows product finetuning under a more relaxed legal framework before it is fully available in the market. However, this logic of a sandbox must be reconciled with the competent authorities' enforcement mandates, particularly when experimentation produces high risks and includes vulnerable populations. Such risks should be minimised or, if possible, excluded *ex ante*. Authorities should be encouraged to have a mandate not only as punitive/repressive enforcement agencies, but also to prevent harm by ensuring voluntary compliance through guidance, instruction and education of potential sandbox participants.

*Baldini and Francis* caution that precisely because sandboxes can entail temporary derogations and discretionary guidance, they risk becoming spaces of "regulated exceptionality" in which fundamental-rights guarantees are relaxed under the banner of innovation.[52] Regulatory sandboxes, they furthermore claim, also risk regulators succumbing to the interests of the industry they are supposed to supervise, rather than protecting the interests of customers and the general public. Moreover, developers may even be in the position to pressure or corrupt officials into prioritising technological innovation over ethical or legal safeguards.[53]

However, while these arguments rightly point to some of the risks of regulatory sandboxes, the conclusion we should draw is not to waive the sandboxes as an intermediary step towards full enrollment of AI tools, but to address these risks. The critique also partly strays away from the core question of how a sandbox should be designed, insofar as they imply deviations from the rule, such as corruption, which can never be excluded. The risk of a regulatory regime being misused should not deter its establishment in the first place. There are strong competing interests which speak in favour of the liability exemption. Regulators who attend the sandbox can acquire technical insights and information on new AI technologies well before they are marketed, enabling them to better calibrate and fine-tune their enforcement activities.[54] Simultaneously, authorities may learn of specific, previously unknown risks and about potential infringements, which they would not be able to enforce against the sandbox participant and others, such as market competitors, as these would fly below their radar. Without a liability exception, the motivation of individual AI developers or even entire market sectors to enter the sandbox regime may completely disappear or significantly diminish. Moreover, it bears noting that AI developers may be drawn to the sandbox regimes perceived as less stringent, if given the option to participate in sandboxes with distinct liability regimes.[55]

[51] *Ahern* [1], p. 1.

[52] *Baldini and Francis* [2], p. 8.

[53] *Baldini and Francis* [2], p. 8.

[54] *Baldini and Francis* [2], p. 6.

[55] EPRS [9], p. 5.

Nevertheless, it is true that the possibility of implementing the liability exemption (Article 57(12) of the AI Act) might be questionable in some cases. In relation to GDPR, the AI Act explicitly stipulates that it does not seek to affect the application of existing EU law governing the processing of personal data, including the tasks and powers of the independent supervisory authorities competent to monitor compliance with those instruments (Recital 10). The liability exemption may thus conflict with rules that do not allow regulators to refrain from imposing a fine despite finding a violation, as is the case under the EU personal data protection legislation.[56] EU data protection legislation applies as soon as personal data is processed, even when it is done in the product development or testing stages, as we can imagine.

This example is particularly illustrative: during testing, the national competent authority observes that the AI system systematically disadvantages candidates based on sensitive attributes such as gender and ethnicity. If, despite guidance from the authority, the developer fails to implement effective mitigation measures, the GDPR equips DPAs with strong corrective tools: they may temporarily or permanently ban the use of the system (Article 58(1)(d) and (f) of the GDPR). In either case, the AI Office must be informed (Article 57(11) of the AI Act), which exposes the potential shortcomings of the participant and may deter sandbox applicants.[57] Since no exemption from liability is possible according to GDPR, the participant cannot rely on the omission of administrative fines in the sandbox, even when they comply with the sandbox plan and follow in good faith the guidance provided by the DPA.[58] According to Article 57(11) of the AI Act, where discretion is built into the relevant law, national competent authorities should use their discretionary powers in a way that supports innovation, which may encourage supervisory authorities to act sensibly where circumstances and national legislation allow for flexibility. However, there are no guarantees, and national and EU authorities have a full remit to supervise and enforce the law against AI sandbox participants.[59]

The AI Act also expressly acknowledges that harms may be inflicted on individuals as a result of the experimentation taking place in the sandbox (Article 57(12)), therefore developers participating in the AI regulatory sandbox remain liable for any damage inflicted on third parties under applicable EU and national liability law. The AI Act makes it clear that AI system providers cannot avoid direct damage claims by users of their tools, even when those are being tested under sandbox conditions. Compensation schemes, such as insurance, could be implemented to limit the risk to participants while protecting affected individuals from potential harm.[60]

[56] In France, Norway and Denmark regulatory sandboxes do not allow participants to derogate from personal data protection law. The French Commission nationale de l'informatique et des libertés (CNIL) even states that it has not launched a 'regulatory sandbox' but only a 'sandbox' as participating in it does not allow the removal of legal constraints (Nathan Genicot and Thiago Guimaraes Moraes, Exploring the boundaries of AI regulatory sandboxes under the AI Act: Flexibility and real-world testing, Cambridge Forum on AI: Law and Governance (2025), 1, e36, p. 3).

[57] *Ahern* [1], p. 33.

[58] *Genicot and Moraes* [11], p. 8.

[59] *Ahern* [1], p. 17.

[60] *Genicot and Moraes* [11], p. 9.

In cases where no liability exceptions apply, the remaining advantage of participating in a regulatory sandbox lies in the close relationship that can be established with regulatory authorities, who can guide participants through the interpretation of the applicable law.[61] Consequently, this alone can result in indirect regulatory compliance, benefiting the participant, regulating authorities and the general public.[62]

## 6 Conclusion

By making at least one national sandbox mandatory in every EU MS by 2 August 2026 (a deadline that might objectively be hard to comply for several MSs) and empowering the European Commission to harmonise participation principles through implementing acts, the AI Act turns what was previously a scattered regulatory technique into a structured component of EU AI governance. Because AI development is intrinsically data-driven, most sandbox projects will involve processing personal data, which must comply with GDPR principles such as lawfulness, purpose limitation, minimisation, storage limitation, transparency, and accountability. Article 59 of the AI Act introduces a tailored framework for further processing of lawfully collected personal data in sandboxes, but only for a narrowly defined set of "substantial public interest" domains and under strict conditions. While the narrow scope of this provision protects the rights of data subjects and should be welcomed, it may exclude socially valuable projects in areas like education and labour, and may potentially incentivise developers to experiment outside sandboxes - where Article 59 safeguards would not apply. Privacy-related risks, such as purpose creep and re-identification, tensions around automated decision-making and meaningful transparency in opaque systems, call for privacy-oriented sandbox design: an obligation to conduct DPIA (as a prerequisite, selection factor, or integrated sandbox requirement), data protection by design and default, and active and coordinated participation of regulatory authorities in a sandbox. With regard to liability exemptions for sandbox participants, the AI Act encourages sandboxes that can strike a credible innovation vs. rights balance, if a mutual understanding and willingness to learn and improve is struck by administrators, participants and regulatory authorities, while effective safeguards against misuse and 'risk-washing' are properly implemented. Accountability under GDPR should remain fully enforceable, but discretion and a sensible (soft?) approach to administrative fining should be encouraged, lest participants be discouraged from joining the sandbox in the first place, if no benefits in terms of rule-relaxation are built into sandbox design.

**Funding information** Funding provided by the Slovenian Research and Innovation Agency (ARIS) under the research programme "Social control, criminal justice system, violence and the prevention of victimizations in the context of high technology market society" (no. P5-0221) and targetted research project "Development of the legal and conceptual framework for the establishment and operation of a national regulatory sandbox for artificial intelligence" (V5-24064), cofunded by the Ministry of Digital Transformation.

[61] *Genicot and Moraes* [11], p. 3.

[62] *Ahern* [1], p. 21.

## Declarations

**Competing interests** The authors declare no competing interests.



## References

1. Ahern, D.: Operationalising AI Regulatory Sandboxes under the EU AI Act: the Triple Challenge of Capacity, Coordination and Attractiveness to Providers. Camb. Forum AI: Law Gov. Forthcoming. https://arxiv.org/pdf/2509.05985
2. Baldini, D., Francis Elizabeth, K.: AI Regulatory Sandboxes between the AI Act and the GDPR: the role of Data Protection as a Corporate Social Responsibility. In: ITASEC 2024: the Italian Conference on Cybersecurity, April 08–11, 2024, Salerno, Italy. https://ceur-ws.org/Vol-3731/paper07.pdf
3. Centre for Information Policy Leadership (CIPL): Regulatory Sandboxes in Data Protection: Constructive Engagement and Innovative Regulation in Practice p. 15 (2019). https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl_white_paper_on_regulatory_sandboxes_in_data_protection_-_constructive_engagement_and_innovative_regulation_in_practice__8_march_2019_.pdf
4. Chałubińska-Jentkiewicz, K., Nowikowska, M.: Artificial Intelligence v. Personal Data, Polish Political Science Yearbook, vol. 51(3) (2022)
5. Commission nationale de l'informatique et des libertés (CNIL) and Information Commissioner's Office – ICO: Regulatory sandbox activities at: https://www.cnil.fr/en/digital-health-and-edtech-cnil-publishes-results-its-first-sandboxes; and https://ico.org.uk/media2/migrated/4019035/sandbox-beta-review.pdf
6. Egbert, S., Leese, M.: Criminal Futures: Predictive Policing and Everyday Police Work. Routledge. https://www.routledge.com/Criminal-Futures-Predictive-Policing-and-Everyday-Police-Work/Egbert-Leese/p/book/9780367643614
7. European Commission: White Paper on Artificial Intelligence, COM (2020) 65 Final (2020). https://commission.europa.eu/document/download/d2ec4039-c5be-423a-81ef-b9e44e79825b_en?filename=commission-white-paper-artificial-intelligence-feb2020_en.pdf
8. European Data Protection Board: Study on the appropriate safeguards under Article 89(1) GDPR for the processing of personal data for scientific research – Final report (2021). https://edpb.europa.eu/system/files/2022-01/legalstudy_on_the_appropriate_safeguards_89.1.pdf
9. European Parliamentary Research Service (EPRS): The impact of the Genereal Data Protection Regulation (GDPR) on Artificial Intelligence. European Parliamentary Research Service (EPRS): the impact of the General Data Protection Regulation (GDPR) on Artificial Intelligence (2020). https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641530/EPRS_STU(2020)641530_EN.pdf
10. Europol: ChatGPT: the impact of Large Language Models on Law Enforcement (2023)
11. Genicot, N., Morae Guimaraes, T.: Exploring the boundaries of AI regulatory sandboxes under the AI Act: Flexibility and real-world testing. Camb. Forum AI: Law Gov. (2025)
12. Jedrzejczak, M.: Protection of Personal Data Processed in Artificial Intelligence Systems (2024). https://researchportal.amu.edu.pl/docstore/download/UAM319db3fdc2e44b818710954f7f02319e/M.+Jedrzejczak,+Protection+of+Personal+Data+Processed+in+Artificial+Intelligence+Systems.pdf?entityType=article&entityId=UAM2746d8d0656d45479386fec4c4b6d426
13. Lyon, D.: Surveillance, Snowden, and Big Data: Capacities, consequences, critique. Big Data Soc. **1**(2) (2014). https://doi.org/10.1177/2053951714541861

14. Mitrou, L.: Data Protection, Artificial Intelligence and Cognitive Services: Is the General Data Protection Regulation (GDPR) Artificial Intelligence-proof? (2018). https://ssrn.com/abstract=3386914
15. Müller, M.: AI-based credit scoring at the intersection of GDPR and AI act: lessons from the SCHUFA judgment. ERA Forum **26**, 647–662 (2025). https://doi.org/10.1007/s12027-025-00865-5
16. Murphy, M.H.: Algorithmic surveillance: the collection conundrum. Int. Rev. Law Comput. Technol. **31**(2), 225–242 (2017)
17. OECD: Regulatory sandboxes in artificial intelligence, OECD Digital Economy Papers, No. 356, OECD Publishing (2023)
18. Sprague, R.: Welcome to the Machine: Privacy and Workplace Implications of Predictive Analytics. Richmond J. Law Technol. **11**(4) (2015). Available at http://jolt.richmond.edu/v21i4/article12.pdf
19. Zarsky, T.: Incompatible: the GDPR in the Age of Big Data (8.8.2017). Seton Hall Law Rev. **47**(2) (2017). Available at SSRN: https://ssrn.com/abstract=3022646
20. Završnik, A.: Criminal justice, artificial intelligence systems, and human rights. ERA Forum **20**(4), 567–583 (2020). https://doi.org/10.1007/s12027-020-00602-0
21. Završnik, A.: Algorithmic justice: Algorithms and big data in criminal justice settings. Eur. J. Criminol. **18**(5) (2021). https://doi.org/10.1177/1477370819876762



**Aleš Završnik** is a Director of the Institute of Criminology at the Faculty of Law in Ljubljana and a Regular Professor at the Faculty of Law, University of Ljubljana. His research interests lie at the intersection of law, crime, technology, and fundamental rights. Završnik collaborated with the Council of Europe's European Commission for the Efficiency of Justice (CEPEJ) on the preparation of the "Ethical Charter on the Use of Artificial Intelligence in Judicial Systems" (2018) and was a member of the Working Group of Experts on AI and Criminal Law established by the Council of Europe's CD-PC. He is an independent Ethics Expert with the European Research Council (ERC). His article "Algorithmic Justice" (European Journal of Criminology, 2021) received the Best Article of the Year Award from the European Society of Criminology.

**Saša Krajnc** is a researcher at the Institute of Criminology at the Faculty of Law, University of Ljubljana. She studied at the Faculty of Law at the University of Maribor, and subsequently worked as a professional associate at the district courts in Ptuj and Ljubljana, and later in the private sector as an expert and consultant in copyright and intellectual property law. She is also a doctoral student at the Faculty of Law at the University of Vienna (Juridicum), currently in the final stages of preparing her doctoral dissertation on copyright in digital environments. Her research focuses on internet law, the enforcement of intellectual property rights online, the regulation of artificial intelligence, and the protection of personal data.