RIVISTA ITALIANA
DI INFORMATICA E DIRITTO
Periodico internazionale di diritto, giustizia e tecnologie

ISSN 2704-7318 • n. 1/2026 • DOI 10.32091/RIID0275• articolo sottoposto a peer review • pubblicato in anteprima il 29 mag. 2026

Davide Baldini

Between the theory and practice of human oversight: Towards specialized AI regulatory sandboxes

Human oversight has become a flagship safeguard in AI governance frameworks, yet its real-world performance is uneven. Empirical studies show that humans supervising automated systems are vulnerable to automation bias and organizational incentives that reduce oversight to a box-ticking exercise. More fundamentally, humans and machines reason in profoundly different ways, making their integration far from straightforward. EU secondary law – most notably the GDPR and the AI Act – requires 'meaningful' and 'effective' human oversight, but offers limited operational guidance on how to achieve it. This paper argues that specialized AI regulatory sandboxes for human oversight can bridge this theory-practice gap. After mapping the legal obligations and practical shortcomings of human oversight, the paper argues how the AI Act's regulatory sandbox regime can be leveraged to test concrete oversight models, metrics, and workflows in controlled conditions. It further provides recommendations for the governance of human oversight-specific sandboxes. The paper thus offers a policy proposal to convert open-textured legislative mandates on human oversight into evidence-based practices, reduce fragmentation across Member States, and strengthen both innovation and fundamental-rights protection in the EU's AI governance.

Human oversight – AI Act – GDPR – Regulatory sandboxes – European Union

Tra teoria e pratica della sorveglianza umana dei sistemi di IA: verso sandbox regolatorie specializzate

La sorveglianza umana dei sistemi di IA ("human oversight") è diventata una delle principali misure di governance dell'IA, ma la sua efficacia pratica risulta disomogenea. Studi empirici mostrano che gli esseri umani chiamati a supervisionare sistemi automatizzati sono vulnerabili all'*automation bias* e a incentivi organizzativi che riducono l'attività di sorveglianza umana a un mero adempimento formale. Ancora più radicalmente, esseri umani e sistemi di IA ragionano in modi profondamente diversi, rendendo l'integrazione tra i due particolarmente complessa. Il diritto dell'Unione – in particolare il GDPR e l'AI Act – richiede forme di human oversight "significative" ed "effettive", offrendo però indicazioni operative limitate su come realizzarle. Questo contributo sostiene che sandbox regolatorie specializzate per la sorveglianza umana dei sistemi di IA possano colmare tale divario tra teoria e pratica. Dopo aver analizzato gli obblighi normativi e le principali criticità applicative della sorveglianza umana, il paper illustra come il regime delle sandbox previsto dall'AI Act possa essere utilizzato per testare modelli di supervisione, metriche e processi operativi in condizioni controllate. Vengono inoltre fornite raccomandazioni sulla governance di sandbox specificamente dedicate alla sorveglianza umana. Il contributo delinea così una proposta di policy volta a convertire obblighi legislativi ampi e indeterminati in pratiche basate sull'evidenza, ridurre la frammentazione tra Stati membri e rafforzare sia l'innovazione, sia la tutela dei diritti fondamentali nella governance UE dell'IA.

Sorveglianza umana – AI Act – GDPR – Sandbox regolatorie – Unione europea

The author is a PhD candidate in European and Transnational Legal Studies, pursuing a double PhD degree between Florence University (Italy) and Maastricht University (the Netherlands)

1. Introduction

Human oversight has long been one of the primary legislative responses to the growing use of AI systems for automated decision-making in our society. Human oversight measures have been embedded in policy discourse as a safeguard against inaccurate and discriminatory algorithmic outputs, as a tool to maintain human accountability over the final decision, and as a symbolic guarantee that critical decisions remain under meaningful human control[1].

In legal and policy terms, human oversight of AI systems comes in different flavours and may take different operational forms. The main ones identified in the literature are the "human-in-the-loop", where the overseer needs to validate each algorithmic output before it can take effect (in which case, the AI system acts as a decision support), "human-on-the-loop", where the overseer has full visibility and power to intervene at any time, for example on request from the affected individual, and "human-in-command", where the overseer retains ultimate decision-making authority over when and how to use the system[2].

Intuitively, human oversight is often framed as offering the best of both worlds: the efficiency, scale, and data-processing capabilities of machines combined with the moral reasoning, contextual understanding, and discretionary judgment of humans. This "Better Together" argument aligns with a long-standing regulatory instinct to counterbalance the perceived dehumanising risks of automation with a reassertion of human agency and is often underpinned by a rhetoric of "trustworthy" and "human-centric" AI, as is notably the case in the context of the EU's Artificial Intelligence Act ("AI Act")[3].

1. For an overview of the widespread adoption of human oversight as a safeguard in many supranational legal instruments related to AI regulation, see: Panezi 2024, pp. 9-11.
2. This tripartition has been endorsed by the influential High-Level Expert Group's Guidelines for Trustworthy AI, which provided the groundwork for the European Commission's first draft of the AI Act (High-Level Expert Group on Artificial Intelligence 2019 p. 16), and subsequently adopted by many AI Act commentators, including: Fink 2025, pp. 2-3; Panezi 2024, p. 12; Enqvist 2023, p. 513.
3. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (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), OJ L, 2024/1689, in particular Rec. (1), (27) and (176).

Yet, as emerging empirical research[4] and insights from cognitive psychology[5] demonstrate, this intuition is often misplaced or, at the very least, overly optimistic. Research reveals that simplistic assumptions about inserting human review into algorithmic decision-making often fail to address its underlying gigantic complexities[6], and may even worsen already-present algorithmic bias and discrimination or introduce unlawful human bias over unbiased algorithmic decisions[7]. In this respect, it emerges that humans do not simply act as neutral "quality-control" agents: they are prone to their own conscious or unconscious biases, misunderstanding how the underlying technology works, often showing over-reliance and under-scrutiny of algorithmic outputs (automation bias), or the inverse problem of algorithmic aversion, when algorithmic outputs are unjustifiably disregarded.

One of the main reasons of these shortcomings is that humans and machines reasoning, despite being apparently similar, function in dramatically different ways; as a consequence, humans and machines are not readily compatible decision-making partners: algorithms mimic human reasoning capabilities, but in fact operate as "prediction machines", operating inferences and producing outputs from extremely vast and obscure (at least, for a human) statistical correlations[8]. This remains true even when they carry out seemingly human-like tasks such as producing natural language outputs, which is the case for AI systems based on Large-Language Models (LLM) that have increasingly been popularized since the launch of ChatGPT in late 2022. Humans, by contrast, do not draw inferences from billions of data points like machines do, but instead rely on qualitative, contextual, and often morally charged reasoning processes[9]. Despite the intuitive appeal of human oversight, the integration of these divergent modes of reasoning is therefore neither straightforward, nor inherently effective.

These structural limits call for regulatory approaches capable not only of imposing abstract oversight duties, but of testing in practice whether said duties can realistically fulfil their intended function. The question, therefore, is not only whether human oversight should exist, but how it can be operationalised in a verifiable and legally coherent manner.

In light of these challenges and given the clear legislative *favor* for human oversight – which has also been openly embraced by the EU legislator in many digital legislative instruments including the

4. *Inter alia*: Gaudeul et al. 2025, which provide compelling evidence that human oversight fails to identify and correct algorithmically generated discrimination when it is not supported by appropriate measures, such as a well-designed human-machine interface, adequate authority and training for the overseer, clear operational guidelines, sufficient time and resources to perform the review.
5. *Inter alia*: Solove–Matsumi 2024, arguing that human and algorithmic reasoning operate according to fundamentally different logics, often making hybrid human and machine decision-making not readily compatible. The analysis demonstrates that human reviewers tend to misinterpret, over-trust, or insufficiently scrutinize algorithmic outputs, meaning that human oversight, if not supported by adequate design, training, and organizational safeguards, fails to correct errors.
6. *Inter alia*: Green 2022, which demonstrates that existing policies mandating human oversight rest on a flawed assumption that humans can reliably detect and correct algorithmic errors. The analysis shows that human-algorithm interaction involves complex cognitive, organizational, and structural constraints; in this context, adding a human reviewer often leaves the underlying problems intact and may even reproduce or exacerbate them rather than mitigate risk.
7. Gaudeul et al. 2025.
8. Wachter 2022.The author gives the compelling example of algorithmically-made statistical inferences that make no sense for a human, but nonetheless correlate with important decisions: "AI also creates seemingly incomprehensible groups defined by parameters that defy human understanding such as pixels in a picture, clicking behavior, electronic signals, or web traffic. These algorithmic groups feed into important automated decisions, such as loan or job applications, that significantly impact people's lives".
9. Solove–Matsumi 2024.

General Data Protection Regulation ("GDPR")[10], AI Act, Digital Services Act ("DSA")[11] and Platform Work Directive ("PWD")[12] – it is necessary to ensure that the concept is not merely rhetorical, but that it meaningfully mitigates the risks associated with automated decision-making, rather than legitimising them. This gap between normative expectations and practical feasibility raises the need for human oversight mechanisms to be experimented with, scrutinised, and refined before being deployed at scale.

It is precisely at this problematic intersection that regulatory sandboxes become relevant. Although usually presented as, primarily, innovation-support tools, sandboxes also serve a second function: they create evidence-generating environments in which regulators and AI operators can jointly test how vague or open-textured legal requirements should work in practice[13]. Because human oversight is one of the most open-textured requirements in EU digital law, and one whose practical shortcomings have been extensively documented, it becomes an ideal candidate for such structured experimentation. This perspective also aligns with the increasing attention that EU digital regulation devotes to experimental governance tools[14], signalling an institutional shift towards more adaptive, evidence-based regulatory approaches to keep up with the pace of technological developments, without hindering innovation.

Originating in the fintech sector and now explicitly embedded in new EU digital regulation including the AI Act, Cyber Resilience Act[15] and Interoperable Europe Act[16] regulatory sandboxes provide controlled environments in which novel technological applications can be tested under the close supervision of competent authorities for a limited time and under defined conditions. Their main purpose is twofold: on the one hand, they allow innovators to experiment with new AI systems while receiving real-time regulatory guidance, thereby fostering legal certainty and supporting compliance; on the other, they enable regulators to gain hands-on knowledge of emerging technologies through a process of "regulatory learning", before such technologies are placed on the market. Such environments are especially useful when laws mandate open-textured or vague obligations, allowing practical testing of how legal requirements might be translated operationally. This seems to be exactly the case of legally mandated requirements of "meaningful" human oversight under the GDPR[17] and "effective" human oversight under the AI Act[18], which remain vague, open-textured, and operationally difficult to implement in light of the inherent issues of human oversight measures. As

10. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 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).
11. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act).
12. Directive (EU) 2024/2831 of the European Parliament and of the Council of 23 October 2024 on improving working conditions in platform work.
13. OECD 2023, p. 13.
14. European Commission 2023.
15. Regulation (EU) 2024/2847 of the European Parliament and of the Council of 23 October 2024 on horizontal cybersecurity requirements for products with digital elements and amending Regulations (EU) No 168/2013 and (EU) 2019/1020 and Directive (EU) 2020/1828 (Cyber Resilience Act), Art. 33(2).
16. Regulation (EU) 2024/903 of the European Parliament and of the Council of 13 March 2024 laying down measures for a high level of public sector interoperability across the Union (Interoperable Europe Act), Art. 11.
17. The "meaningfulness" requirement for human oversight is not explicitly set forth in the GDPR text, but has been carved out by the Article 29 Working Party as a minimum threshold to ascertain whether a decision is based solely on automated processing under the meaning of Art. 22(1) GDPR. See Article 29 Working Party 2017. This concept of "meaningful human oversight" has then been widely adopted by case-law, as noted by Lazcoz–De Hert 2022.
18. More precisely, Art. 14(1) AI Act requires that high-risk AI systems "can be effectively overseen by natural persons during the period in which they are in use".

they offer a structured setting for supervised experimentation, sandboxes may hold the potential to bridge the gap between the law's mandatory safeguards on oversight, and the complex, challenging realities of supervising algorithmic decision-making processes in practice, thereby contributing to the development of clearer practices, more effective oversight mechanisms, and ultimately more trustworthy AI governance.

In light of the above, this paper seeks to contribute to the current academic and policy debate around both human oversight and regulatory sandboxes in EU law, by introducing and advancing the concept of specialized regulatory sandboxes as a means to enable competent authorities and AI operators to design and implement effective and meaningful human oversight in an evidence-based way. We depart from a general-purpose and even a sector-based sandbox regime, by proposing the establishment of "requirement-centric" sandboxes, focused on human oversight requirements under the AI Act and connected EU digital legislation such as GDPR's Article 22 and PWD's Article 10. Unlike existing vertical and sector-specific regulatory sandboxes (e.g., for automotive, healthcare, etc.), our model centres on the oversight requirement itself.

To do so, the paper is structured in five parts. While Section 1 provides the introduction and sets the stage for the following sections, Section 2 situates human oversight in the broader international landscape and then details the EU framework, focusing on the GDPR and the AI Act overlapping requirements, with a view to clarify the respective roles of providers and deployers (also as data controllers under the GDPR). Section 3 reviews the empirical and theoretical work on the practical shortcomings of human oversight – cognitive, organizational, and systemic. Section 4 introduces regulatory sandboxes: it defines the instrument, traces its emergence, and sets out the EU legal basis and design under the AI Act, highlighting their potential for achieving regulatory learning and advancing legal certainty in the context of human oversight requirements. Section 5 advances the core proposal for specialized AI regulatory sandboxes for human oversight, also sketching some high-level policy advice on how to operationalize them. The conclusions in Section 6 present the main findings and propose a path for further research and policy development.

2. Human oversight obligations under EU law

Human oversight has become a common feature across the main international instruments developed in multilateral forums to regulate or guide the governance of artificial intelligence at both global and regional levels[19]. Most notably, the Council of Europe's Framework Convention on AI[20], the UN General Assembly Resolution on AI adopted on 21 March 2024[21], and UNESCO's Recommendation on the Ethics of Artificial Intelligence of 23 November 2021[22] all enshrine explicit provisions on human oversight. The influential OECD Recommendation on AI, first adopted on 21 May 2019 and amended on 4 May 2024, also refers to the need for human agency and oversight[23].

In line with this worldwide trend, one of the defining features of EU digital regulation is its insistence on having humans oversee the functioning of algorithmic systems. From data protection to artificial intelligence governance, the EU's approach is explicitly human-centric[24], aiming to ensure that rapid technological advances remain aligned with fundamental rights[25]. This focus on human oversight is evident in cornerstone EU digital laws like the GDPR and – as anticipated and even more so – the AI Act. Both instruments boast a large,

19. Further regional or national initiatives are further analysed by Panezi 2024, pp. 9-11.
20. Council of Europe 2024. The Convention is the first binding international treaty on AI regulation.
21. United Nations 2024.
22. UNESCO 2021.
23. OECD 2024.
24. This concept is explicitly referred to in the AI Act as a cornerstone of the EU approach to AI regulation. See: Rec. (1), (6), (8), (27), (176) and Art. 1.
25. Fink 2025.

horizontal scope of application[26], and embed requirements for human supervision of automated processes if certain thresholds are met[27] – a legal safeguard reflecting the conviction that even in an age of AI-enabled automated-decision making, critical decisions should not be left entirely to machines. Other, narrower digital regulations, notably including the PWD (Art. 10), also mandate overlapping human oversight requirements, adding weight to the EU's legislator favourable approach to human oversight as one of the main remedies against algorithmic harms. By requiring a human hand on the controls, EU law seeks to preserve accountability, prevent harm to fundamental rights, and ensure trust in digital technologies[28].

2.1. Human oversight in the GDPR

The GDPR, directly applicable since 25 May 2018, was among the first laws worldwide to recognize, in its Article 22, both a prohibition[29] of fully automated decisions that produce legal or similar effects (subject to some exceptions), and a right to contest the fully automated decision, when such exceptions are met. This provision equips individuals with a *prima facie* right not to be subject to a decision based solely on automated processing, if said decision produces legal or similarly significant effects over them, thus expressly requiring a human-in-the-loop[30]. Even when such automated decisions are exceptionally allowed under paragraph 2 of the Article (i.e., based on data subject's explicit consent, contractual necessity, Union or Member State legal provision), the controller must provide "suitable measures to safeguard the data subject's rights and freedoms and legitimate interests" including as a minimum the right to obtain human intervention, to express their viewpoint, and to contest the decision, thereby producing an "on request" human-on-the-loop obligation[31]. In other words, whenever algorithms make important determinations about people, the GDPR mandates that a human can step in to review (before it takes effect) or override (after it has taken effect) the outcome. Recital 71 underscores the rationale: it notes that purely automated decisions need to include mechanisms for human oversight and the ability for individuals to challenge the outcomes on request, via a right to human intervention[32]. This was a landmark statement of principle in data protection law – and one tracing its roots to the already influential Directive 95/46/EC[33] –, as it expressly bridged data protection and automated decision-making regulation, showcasing the EU's early legislative response to concerns of "black box" algorithms producing inscrutable and unaccountable decisions. By requiring "meaningful" human

26. Although the AI Act decouples its (very large) formal scope of application from its (very narrow) material scope of application: broad notion of "AI system" is leveraged to establish the law's boundaries, but actual obligations are then placed on a much smaller subset of AI systems (mostly, high-risk ones), as critically noted by Veale–Zuiderveen Borgesius 2021, pp. 108-110.
27. Such thresholds being, in the case of the GDPR, having automated systems take decisions with legal or similarly significant effects on individuals under Art. 22, while, in the case of the AI Act, having AI systems qualifying for the "high-risk" category under Art. 6.
28. Fink 2025; High-Level Expert Group on Artificial Intelligence 2019; See also Rec. (27) of the AI Act, which strongly underlines the link between the AI Act and the principles developed by the High-Level Expert Group on AI.
29. The text of Art. 22(1) GDPR, is worded in positive terms, rather than as a prohibition; however, it was interpreted as such firstly by the EU data protection authorities in 2017 (Article 29 Working Party 2017, pp. 19-20), then endorsed by the Court of Justice (Court of Justice, *Land Hessen/SCHUFA Holding AG*, case C-634/21, p. 52), despite some resistance in the literature (Tosoni 2021).
30. Lazcoz–De Hert 2022, pp. 12-14.
31. Art. 22(3) GDPR. Lazcoz–De Hert 2022, pp. 12-15. The authors characterize this obligation as a type of "human-out-of-the-loop" measure, in contrast with the "human-in-the-loop" requirement in Art. 22(1) GDPR.
32. Dirutigliano–Baldini 2025. Rec. (71) is also notable for expressly linking Art. 22 GDPR, to non-discrimination objectives, requiring data controllers to mitigate the risk of inaccuracies and discriminatory outputs.
33. Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data.

review, the GDPR aims at embedding fairness, redress mechanisms, and – ultimately – dignity, namely the values that human judgment can intuitively provide[34].

The practical application of the provision has cemented this approach: Article 22 GDPR has been extensively interpreted by both the Member States' courts and data protection authorities at the outset of GDPR applicability[35] and, subsequently, further expanded by the Court of Justice of the EU by leveraging the fundamental rights-based foundation of the GDPR in the treaties[36]. These favourable interpretative developments of the provision have, in turn, spurred numerous concrete applications of Article 22 by Member States' supervisory authorities and national courts[37] – with a particular relevance in the labour law domain, where it has been strategically leveraged by workers' associations to challenge algorithmic surveillance practices by employers[38] – significantly influencing practices around automated decision-making, both within the EU and internationally, thanks to the EU's regulatory influence in the digital domain.

2.2. Human oversight in the AI Act

Building and expanding on this foundation, the AI Act considers human oversight as an explicit obligation for high-risk AI systems. The central provision in this respect is Article 14, aptly titled "Human oversight". This Article places obligations aimed at providers, in line with the AI Act's overarching nature as a product-based legislation[39], essentially requiring that all high-risk AI systems – meaning, those AI systems that fulfil the requirements laid down in Article 6 of the AI Act – be designed and developed so that "they can be effectively overseen by natural persons" during their use, by means of appropriate safeguards such as human-machine interfaces, and by illustrating the chosen safeguards in the system's instructions[40]. Deployers of AI systems are then required to assign human oversight in practice, by making use of the aforementioned providers' safeguards, and in line with the provider's instructions, which must also indicate the measures aimed at enabling effective oversight[41].

According to the letter of the provision, human oversight measures to be envisaged by providers must aim at preventing or minimising "risks to health, safety, and fundamental rights" that might emerge from the system's operation, in particular where such risks persist despite the application of other requirements aimed at high-risk AI system providers[42].This seemingly entails an inverse relationship between human oversight and other obligations aimed at high-risk AI system providers[43], tracing its origin to the work of the High-Level Expert Group on AI[44], which set the basis for the AI Act. This inverse relationship conceptualizes human oversight as a "safety net" within the broader context of the AI Act, namely, a safeguard whose significance grows as other risk management mechanisms prove insufficient or fail to adequately mitigate the risks posed by a given AI system.

34. Although, as shall be seen in Section 3, *infra*, the actual effectiveness of human oversight to achieve these aims is – to say the least – heavily debated.
35. For a summary of national cases concerning the application of Art. 22 GDPR, see: Barros Vale–Zanfir-Fortuna 2022; for the interpretation adopted by EU data protection authorities, see: Article 29 Working Party 2017.
36. Court of Justice, *Land Hessen/SCHUFA Holding AG*, case C-634/21.
37. Barros Vale–Zanfir-Fortuna 2022.
38. Cansu–Farrar 2021, which showcase many instances where data protection legislation – and especially Art. 22 GDPR – has been leveraged by platform workers and their representatives to challenge unfair algorithmic management practices by gig economy platforms.
39. Veale–Zuiderveen Borgesius 2021.
40. Art. 13(3)(d) AI Act.
41. See Art. 26, para. 1 and 2 AI Act.
42. Art. 14(2) AI Act.
43. This elevates human oversight as a key safeguard in the overall context of the AI Act, as noted by Fink 2025, and Enqvist 2023.
44. High-Level Expert Group on Artificial Intelligence 2019.

This arrangement goes to show the high trust that the EU legislator has put on human oversight as a safeguard against algorithmic harm, elevating it as a key component of the EU's human-centric approach to AI, in line with the High-Level Expert Group on AI's position[45].

As noted, the bulk of the legal obligations are placed on the provider, to whom Article 14 is entirely addressed. However, in doing so, and in contrast to other obligations, the provision leaves much flexibility to the provider. Article 14(3) AI Act leaves to the latter the choice between integrating built-in oversight mechanisms into the AI system or specifying measures to be implemented by the deployer, or a combination of both, depending on the specific risks associated with the AI system's intended use.

Furthermore, and contrary to Article 22 GDPR which does not expressly prescribe oversight measures for data controllers, Article 14(4) AI Act lists five specific human oversight measures, spanning from the obligation to develop the AI system in a way that enable humans to interpret its outputs, or to remain aware of automation bias, to a built-in "stop" button that allows the overseer to suspend the system use. However, much room for flexibility is present also in this case, as the provider is required to choose and implement such measures "as appropriate and proportionate". As noted in the literature, whether this flexibility extends to "which" measures to implement or remains limited to the "how" is unclear, although a teleological interpretation may suggest the latter option[46].

Finally, an enhanced human oversight regime in the form of a human(s)-in-the-loop requirement is established for remote biometric identification systems, where identification of a natural person by the system must be separately verified and confirmed by at least two natural persons[47], subject to targeted exemptions for AI systems used in law enforcement and migration.

Although Article 14 remains the central provision, the human oversight framework is completed by two other provisions that are in a strict relation with it, and that contribute to extend human oversight obligations to the deployer.

Firstly, Article 13 – titled "Transparency and provision of information to deployers" – requires the provider to include within the AI system's instructions for use the human oversight measures adopted pursuant to Article 14, including the technical measures to ensure the system's interpretability[48]. Lastly, as already recalled, Article 26 completes the framework by mandating deployers to use the system in accordance with said instructions, as well as to specifically "assign human oversight to natural persons who have the necessary competence, training and authority, as well as the necessary support"[49].

In a nutshell, while the provider must equip the system to enable human oversight, the deployer is then required to make full use of the oversight functionalities. The legal design around human oversight in the AI Act seems sound and well-structured: by "hard-wiring" human oversight into the high-risk AI system lifecycle, from its design phase (where the provider is the main actor) all over to the deployment phase (where the deployer takes the lead), the EU legislator aims to avoid scenarios where algorithmic tools operate unchecked and to reassure the public that accountable humans remain ultimately in charge of critical decisions. Nevertheless, many issues loom over the effective achievement of the EU's stated goals regarding human oversight, as will be discussed in the next section.

3. The practical shortcomings of human oversight

However well-crafted the EU's legal provisions on human oversight are, a growing body of research suggests that effective oversight is often more aspirational than real, and that human oversight measures – even when they carefully follow the letter of law – can sometimes even exacerbate existing issues. One recent study, for instance, found that humans tasked with revising AI systems' outputs

45. *Ibidem* p. 16.
46. Enqvist 2023 p. 10.
47. Art. 14(5) and Rec. (73).
48. Art. 13(3)(d).
49. Art. 26(2).

in the context of hiring recommendations added their own "human-made" bias on top of discriminatory outputs produced by the AI system[50], or even introduced *ex novo* bias and discrimination where the AI system had not done so[51]: for example, judges in several jurisdictions were found to override algorithmic risk assessments in ways that increased racial disparities, compared to the original algorithmic output[52]. Such findings highlight that simply having a human in the loop is no panacea: the human can fail to catch the algorithm's mistakes or can even aggravate them.

3.1. Cognitive and psychological limits

There are a few reasons for this. Cognitive biases pose a first major hurdle. Human decision-makers tend to place undue trust in algorithmic outputs – a phenomenon known as "automation bias", where people defer to automated recommendations and apply less independent scrutiny than they otherwise would. Research shows that even trained personnel tasked with reviewing machine suggestions can become passive, failing to double-check or countermand the algorithm's output[53]. On the other side of the spectrum lies "algorithmic aversion" where individuals unjustifiably reject AI-generated outputs, potentially replacing fair and consistent algorithmic decisions with discretionary judgments tainted by human error or unlawful bias[54].

The fundamental issue is that, although AI outputs may closely resemble human logic, they are derived from statistical inferences, which differ significantly from moral and contextual judgments taken by humans[55]. Although AI systems – including generative AI systems – may appear to closely replicate human reasoning, in reality they function primarily as predictive tools, generating outputs based on statistical inferences and correlations drawn from vast amounts of data. As a result of this stark diversity, humans often perform badly when reviewing algorithmic outputs: in practice, the human overseer may exhibit only nominal vigilance, typically by rubber-stamping the algorithm's decision due to over-confidence in the machine's capabilities, responsibility avoidance, lack of training, resources or time pressure. Notably, empirical studies find that people struggle to judge when an AI's recommendation is erroneous, often overestimating the accuracy of the system, or ignoring contradictory cues[56]. Depending on the features of the AI system, this can lead to omission errors (failing to act because the system gave no alarm) and commission errors (uncritically following a flawed recommendation). Crucially, although Article 14 of the AI Act explicitly acknowledges automation bias as an issue to be addressed by the provider[57], inherent human cognitive limitations cannot be easily "legislated away": despite the express legal obligation in the AI Act to account for automation bias, the way in which this problem is to be addressed in practice is not fully clear.

The complexity and opaqueness of advanced AI systems present further practical limitations to meaningful oversight. AI systems often operate as complex black-box models, making it exceedingly difficult for human overseers to understand the rationale behind a given output. The AI Act also addresses this issue, by requiring that providers design systems to be "effectively overseen" by humans[58], and that the provider's instructions regarding human oversight include explanations of technical measures to make algorithmic outputs interpretable to humans[59]. However, while this assumes that explanations or transparency measures

50. For an overview of the challenges faced by the EU anti-discrimination law framework vis-à-vis the rise of automated decision-making, see Adinolfi 2022.
51. Gaudeul et al. 2025.
52. Green 2022.
53. Gaudeul et al. 2025.
54. *Ibidem*.
55. Solove–Matsumi 2024.
56. Gaudeul et al. 2025.
57. Art. 14(4)(b) AI Act.
58. Art. 14(1) AI Act.
59. Art. 13(3)(d) AI Act.

will enable comprehension, evidence casts doubt on that assumption: studies show that providing algorithmic explanations or other transparency tools does not necessarily improve an operator's ability to detect errors or biases[60]. Ironically, interpretability and explanations can instead have the perverse effect of even increasing human trust in the AI system's decisions – leading oversight personnel to become more confident in the algorithmic outputs, even when they are wrong or discriminatory.

3.2. Organizational and systemic constraints

Furthermore, and perhaps most troubling, are the organizational and systemic shortcomings that can pressure human overseers into rendering their activity a purely symbolic exercise. Despite frequent policy claims that human oversight redresses algorithmic harms, empirical support for such claims remains limited and often anecdotal[61]. Conversely, empirical research overwhelmingly indicates that humans cannot reliably perform the safeguarding functions envisioned by law[62]. In practice, human oversight frequently provides only the appearance of control – what some scholars have labelled a "skeuomorphic" humanity[63] – giving the impression of a human choice, but without its substance. The human overseer may technically have discretion to override or alter an automated decision, but organizational norms and the design of the workflow often discourage deviation: most commonly, the human overseer is nudged into passively accepting the algorithmic output, to avoid assuming responsibility in cases where the decision proves wrong or harmful in practice. As a result, oversight can deteriorate into a merely formal requirement, where the human's role is nudged into confirming the machine's output rather than critically reviewing it.

Ultimately, existing research warns that current reliance on human oversight as a safeguard is not delivering on its intuitive promise to remedy algorithmic injustice. On the contrary, oversight measures may legitimize the deployment of flawed algorithms by cloaking them in a veneer of human review, all the while shifting responsibility for any harm away from providers, and onto lower-level human operators, acting as "liability sponges"[64].

The EU's ambitious framework thus confronts a paradox: it elevates human oversight as one of the central pillars to safeguard fundamental rights when operating high-risk AI systems, yet cognitive barriers, real-world human frailties and systemic pressures all conspire together to prevent that oversight from being genuinely effective. This tension sets the stage for exploring alternative approaches – a task the following section will undertake – but it reinforces a critical point: without acknowledging and addressing these practical shortcomings, human oversight risks remaining an attractive principle on paper that, in the best case, delivers only modest safeguards in practice, while in the worst case exacerbates algorithmic harms.

The considerable issues and shortcomings explored above do not mean, however, that human oversight is bound to remain an ineffective safeguard. Furthermore, the fact remains that EU secondary law explicitly requires AI system operators to implement oversight in a manner that is both meaningful and effective, as illustrated in the previous section. This situation implies that when establishing oversight measures, operators must adequately consider and address the shortcomings discussed above. This requirement applies both within the AI Act framework, where obligations are allocated between providers and deployers, and under the GDPR, where obligations rest on the data controller and are underpinned by the foundational principle of accountability[65], albeit with different level of intensity,

60. Gaudeul et al. 2025.

61. Green 2022, p. 2.

62. *Ibidem*.

63. Brennan-Marquez–Susser–Levy 2019.

64. Green–Kak 2021.

65. Arts. 5(2) and 24 GDPR. Applied in the case at hand, this principle arguably mandates data controllers to identify and implement measures aimed at ensuring meaningful human oversight is achieved in practice, taking into account and solving its relevant shortcomings.

as discussed in the previous paragraph. In this respect, it is essential to recall that the AI Act and GDPR typically operate concurrently when high-risk systems are developed or deployed, as processing of personal data is almost invariably present in these types of systems.

This combination of a strong and horizontal legal mandate, on the one hand, and persistent practical and interpretative indeterminacy, on the other, calls for regulatory instruments capable of producing empirical and operational knowledge on how legally-mandated human oversight obligations can be meaningfully implemented in practice. In the absence of such instruments, there is a tangible risk that human oversight remains either under-specified in abstract guidance or over-formalised in compliance practices, without any genuine benefit in addressing the risks it seeks to mitigate.

It is precisely at this junction between normative ambition and operational uncertainty that regulatory sandboxes acquire particular relevance. By design, sandboxes provide a structured and supervised setting in which legally mandated requirements can be experimented with in concrete technical and organisational contexts, allowing regulators and AI operators to jointly observe how those requirements function in practice, where they fall short, and under which conditions they may be rendered effective. As such, sandboxes are uniquely positioned to translate legal obligations on human oversight into evidence-based practices that can later inform guidance, standard-setting and enforcement, before such obligations are generalised and applied at scale.

4. AI regulatory sandboxes under the AI Act

4.1. Origins and rationale of regulatory sandboxes

Regulatory sandboxes are an innovative legislative experimentation tool which has emerged during the last decade as a response one of the main issues inherent to the regulation of new technologies, namely the so-called "pace problem" where technological innovation evolves significantly faster than the ability of legislators and regulators to adapt existing legal frameworks[66]. This challenge is further compounded by the "Collingridge dilemma"[67], which highlights the difficulty of regulating emerging technologies early on – when intervention is easier, but knowledge is limited – and the corresponding difficulty of intervening later, when the consequences are better understood, but the technology has become entrenched and resistant to change[68].

In this respect, regulatory sandboxes have emerged as a promising solution in many jurisdictions. No commonly shared definition of this instrument exist[69], but it can be broadly framed as supervised testing spaces that allow organizations to trial new and innovative technological products, overseen by regulators and restricted to a set period of time[70].

Beyond definitional uncertainties, regulatory sandboxes may take two main forms regarding their scope of operation: horizontal and vertical sandboxes[71]. On the one hand, the former (also called "cross-sector" or "cross-domain" sandboxes), are general-purpose environments open to a wide range of technologies and industrial sectors[72]; they

66. Bagni–Seferi 2025, p. 19; Downes 2009.

67. Collingridge 1980.

68. Moraes 2023, pp. 2-4.

69. Cirone 2025, p. 262, where the author observes that a plethora of different definitions are given by policy documents and relevant sectoral legislations that establish regulatory sandboxes; ultimately, the definition varies depending on the context of use.

70. *Ex multis*, a similar definition is provided by Madiega–Van De Pol 2022.

71. Lanamäki et al. 2025, p. 42, noting that the AI Act implicitly allows both domain-specific and horizontal sandboxes. Ranchordas–Vinci 2024, pp. 9-11; Longo 2021.

72. In the Italian context, a salient example is the regulatory sandbox named "*Sperimentazione Italia*" ("Italy experimentation"), set forth by Art. 36 of Law Decree No. 76/2020, as converted by Law No. 120/2020. For a comment, see: Ranchordas–Vinci 2024.

aim to facilitate innovation broadly and to support regulatory learning across diverse use cases, but their breadth inherently limits the depth of experimentation on any single requirement. On the other hand, vertical sandboxes (also called "sector-specific" or "domain-specific" sandboxes), focus on a defined domains such as healthcare, connected vehicles or public administration[73]. This specialisation allows regulators to develop sector-specific expertise through regulatory learning, and to generate more granular and operational guidance, as illustrated by several national experiences[74].

First pioneered in the fintech sector by the UK Financial Conduct Authority's sandbox in 2015[75], as noted above, regulatory sandboxes promise to bridge the longstanding gap between rapid technological evolution, on the one hand, and the inherently slower pace of traditional legislative processes and enforcement, on the other. One of the primary advantages of regulatory sandboxes is their capacity to allow both regulators and innovators to examine and better understand innovative technological applications and their risks, before these enter the market: a process known as "regulatory learning"[76]. By enabling a proactive, *ex-ante* scrutiny, sandboxes significantly reduce the risk of regulatory approaches rapidly becoming obsolete or inadequate in the face of emerging technological applications, allowing regulators to adjust their interpretation of existing norms, or even prompt the legislator to revise the regulatory framework itself. On the other hand, organizations that attend the sandbox benefit from the possibility to develop and test their innovative solutions with real-time guidance from regulators, thus facilitating legal certainty, easing compliance requirements and accelerating market entry[77].

4.2. Design and operation in the EU context

Due to these and other benefits, as of late sandboxes have seen widespread global adoption[78] across diverse technological sectors – extending far beyond their original fintech origins – as flexible tools that effectively balance technological advancement with responsible regulatory oversight. The EU makes no exception, as it has followed this global trend and incorporated this instrument into some of its recent digital law acquis, including the AI Act[79], which explicitly foresees "AI regulatory sandboxes" within its Chapter VI ("Measures in support of innovation"), indirectly defining them in Article 57 as supervised frameworks that "provide a controlled environment that facilitates the development, testing and validation of innovative AI systems for a limited time before their placement on the market"[80]. These provisions anchor sandboxes in EU law as a means to spur AI innovation in a compliant, well-monitored setting. Eligibility for sandbox participation is open to AI providers (particularly, startups and SMEs, who are given priority access) that meet criteria that are still to be detailed in EU Commission's implementing acts in accordance with Article 58.

73. Still in the Italian context, an example is the Italian FinTech Regulatory Sandbox, established by Art. 36 of Law Decree No. 34/2019, converted into Law No. 58/2019. Ranchordas–Vinci 2024, pp. 10-11.
74. One example is the French Data Protection Authority's sandbox, further explored in section 5, *infra*. See also Bagni 2023.
75. OECD 2023, p. 15.
76. Ranchordas–Vinci 2024; Bagni 2023.
77. Bagni–Seferi 2025.
78. For example, as noted by Madiega–Van De Pol 2022, Japan introduced in 2018 a regulatory sandbox framework open to organizations and companies, both domestic and international, interested in experimenting with emerging technologies such as blockchain, artificial intelligence (AI), and the Internet of Things (IoT). This framework applies across various sectors, including financial services, healthcare, and transportation.
79. This concept is not limited to AI: the EU Cyber Resilience Act, in its Art. 33(2), likewise envisages cybersecurity sandboxes, allowing novel digital products to be developed and tested under regulatory supervision for a limited time before market entry (with an emphasis on supporting SMEs). However, while each Member State is required to establish at least one regulatory sandbox under the AI Act, the Cyber Resilience Act leaves the establishment of cybersecurity regulatory sandboxes as a mere possibility.
80. Art. 57(5) AI Act.

Under the AI Act, each sandbox operates pursuant to a specific testing plan agreed between the high-risk AI system providers, as participants, and the national competent authority, and is time-bound, with extensions possible based on project needs. The sandbox's supervisory structure entrusts national market surveillance authorities to oversee the experiments, in coordination with other relevant bodies where necessary. Most notably, Article 57(1) expressly requires that the competent national Data Protection Authority ("DPA") be involved in case where AI systems involve the processing of personal data, which is bound to happen in most cases given that high-risk AI systems envisaged under the AI Act often involve the processing of personal data in the context of their training, testing and/or operation[81].

Importantly, involved authorities retain full oversight powers: participation in a sandbox does not exempt AI operators from fundamental legal requirements or from civil liability in case it arises, but administrative fines cannot be imposed in case of *bona fide* failure to comply with the law. Moreover, participating authorities are required to provide guidance on regulatory expectations and are expected to exercise their oversight in a flexible, innovation-friendly manner.

In addition, the AI Act has foreseen further incentives for attending an AI regulatory sandbox, most notably by producing legal effects aimed at encouraging participation: most notably, once a specific sandbox experiment has been successfully attended, the competent authority is required to provide an "exit report", which providers may then leverage to demonstrate their compliance with the AI Act[82].

In this context, AI regulatory sandboxes seem intuitively well-suited to solve the practical problems posed by human oversight, discussed in Section 3, above. In light of their features, sandboxes are arguably the main tool to achieve one of the AI Act's stated aims, namely, supporting innovation[83]: from this perspective, they act as a pragmatic regulatory learning instrument, supporting an evidence-based regulatory approach to addressing ambiguous, problematic and operationally challenging requirements, which is precisely the case for human oversight requirements for high-risk AI systems. Through dialogue with relevant AI Act and GDPR enforcers – which are required to provide feedback on regulatory expectations to participants – and testing within these controlled environments, sandboxes may help to operationally clarify many of the legislative requirements in practice.

In this respect, the express obligation to involve Data Protection Authorities in the operation and supervision of sandboxes appears crucial. DPAs are in this case mandated to provide interpretative guidance on GDPR requirements, together with the national AI Act's Market Surveillance Authority[84]. The joint participation of GDPR and AI Act enforcement authorities is particularly significant, as it creates an unprecedented institutional forum for addressing cross-cutting compliance challenges in real time. If successful, the joint operation of sandboxes may help to inform new, evidence-based interpretations of the law, with specific regard to legislative overlaps between the AI Act and the GDPR. This is especially relevant in the case of human oversight, as the requirement simultaneously serves as a safeguard under Article 22 GDPR and as a design and deployment obligation under Article 14 AI Act, thereby making coordinated guidance a *de facto* mandatory element to ensure coherent and practicable implementation of both laws.

81. This requirement gives rise to a number of interpretative issues in light of the overlap between GDPR and AI Act provisions on the processing of personal data, particularly Art. 59 of the AI Act. For an analysis, see *inter alia*: Bagni–Seferi 2025, pp. 70-84, concluding that the stringent conditions established by Art. 59, as well as its cumulative applications with other GDPR and AI Act provisions on data governance, may invertedly discourage personal data re-use in the sandbox; Holtz–Ledendal 2026, pp. 11-12, highlight instead that Art. 59 even provides a legal basis to process special category data pursuant to Art. 9(2)(g) GDPR.

82. Art. 57(7) AI Act.

83. Art. 1(1) AI Act.

84. It should be noted, depending on the Member State's choice, the Data Protection Authority and Market Surveillance Authority may coincide on the same entity (Art. 70(1) AI Act).

Despite these potential benefits, some challenges stand in the way. Regulatory sandboxes foreshadow the possibility of regulatory capture[85], they may inadvertently produce regulatory fragmentation at the national or even regional and local level in case where the results of sandboxes are not sufficiently publicized and generalized. Furthermore, they require extensive resources, and expertise from both the attendees and the regulators, in order to function properly. Despite this, the clear legislative endorsement at the EU level, combined with heavy investments already foreseen under initiatives such as the Digital Europe Programme and related Coordination and Support Actions, suggests that these practical barriers may soon be addressed, paving the way for more robust, specialized regulatory sandboxes under the AI Act. These problems and relevant solutions have been extensively explored and are acknowledged by the EU legislator, which has envisaged various tools to deal with them[86].

In sum, by closing the gap between high-level legal mandates and operational realities, regulatory sandboxes hold the promise to mitigate interpretative uncertainties and leave open the possibility for more targeted initiatives. This leaves the door open to the establishment of one or more specialized "human oversight sandboxes", a proposal to be addressed at a high-level in the following section.

5. A proposal for specialized AI regulatory sandboxes for human oversight

To optimize their effectiveness and to help them achieve their aims, this paper suggests that AI regulatory sandboxes should be organized around thematic streams, while remaining at the same time aligned with the flexible legal requirements and safeguards of the AI Act's sandbox regime, outlined in its Chapter 6. It argues, in particular, to go beyond the traditional "horizontal versus vertical sandboxes" dichotomy, so far limited to distinctions between broad cross-sector and industry-specific sandboxes; instead, it suggests advancing a model of vertical specialization grounded in AI-specific regulatory requirements, in this case devoted explicitly to testing and refining human oversight mechanisms for high-risk AI systems.

While regulatory sandboxes traditionally take as their starting point a specific product or service and then derive the applicable legal obligations, the model advanced in this paper deliberately reverses this logic. Here, the analytical anchor is not the technological application, but a single, cross-cutting regulatory requirement of the EU digital acquis, that is, human oversight. This inversion reflects the particular nature of human oversight obligations, and their relevant shortcomings, as discussed in Sections 2 and 3.

For this reason, the traditional product-based approach inherent in most regulatory sandboxes is, arguably, insufficient. The feasibility and effectiveness of human oversight depend on a constellation of factors – technical design choices, organisational arrangements, user capabilities, contextual constraints, and the interplay between multiple legal regimes (most notably: AI Act, GDPR, PWD). It is suggested that these elements cannot be meaningfully captured by examining all requirements applicable to an AI system. A requirement-centric sandbox, by contrast, would allow regulators to isolate the oversight obligation itself, observe its operationalisation across heterogeneous technological and organisational environments, and identify recurring patterns that a product-centric experiment is structurally unable to reveal. In this sense, reversing the traditional sandbox logic is not only justified; it is necessary to generate evidence-based, generalisable guidance on a horizontal obligation whose proper implementation is central to the EU's human-centric approach to AI governance.

This shift in perspective should also be positioned in the context of existing experimentation practices in the EU. While no Member State has yet deployed a sandbox explicitly centred on specific requirements such as human oversight, some EU Member States have already conducted data protection-related sandboxes on specific sectors,: one recent example is the French Data Protection Authority's sandbox frameworks, which focused sequentially on healthcare, education, and public administration[87]. These initiatives have already yielded promising results despite not being yet able

85. Bagni–Seferi 2025, pp. 192-206.

86. For example: *Ibidem*; Ranchordas–Vinci 2024.

87. Commission Nationale de l'Informatique et des Libertés 2025.

to rely on relevant EU law provisions, such as those included in the AI Act, and thus enjoying very limited regulatory flexibility[88].

In this respect, one notable and recent illustration comes from the abovementioned French Data Protection Authority's sandbox framework on public administration[89]. In particular, the Authority examined *Conseils Personnalisés*, an AI-based recommendation tool deployed within the *France Travail*, the French public employment services. The system, built around a locally-hosted large language model enhanced through Retrieval-Augmented Generation (RAG), suggested tailored training opportunities to jobseekers based on their profile and on inputs provided by employment advisers. Although the tool did not directly take binding decisions, it shaped the options that advisers present to users and thus raised questions linked to Article 22 GDPR, especially in light of the recent Court of Justice of the EU's *SCHUFA*[90] case, and to the threshold above which human intervention could be deemed "meaningful"[91]. The most delicate issue concerned the requirement that decisions must not be based solely on automated processing, under Article 22(1) GDPR, thereby mandating the presence of a human-in-the-loop mechanism[92]. In this respect, while some measures were already put in place by the data controller, such as internal guidelines, training modules, and appropriate working conditions ensuring that no pressure was placed on advisers to follow the system's recommendations, the sandbox experiment identified further measure to ensure that the "meaningfulness" requirement was achieved: *France Travail* developed an AI-awareness pathway including continuous training, specific internal support for the use of the *Conseils Personnalisés* tool, and the appointment of "AI correspondents" acting as intermediaries between local agencies and the central administration. Under these conditions, as identified during the course of the sandbox, the Authority concluded that proposing training opportunities via the tool did does not appear to constitute fully automated processing[93].

Drawing from this and other experiences at national level[94], scholars have recently highlighted the practical benefits of sector-specific approaches to sandbox design, emphasizing that such targeted arrangements allow deeper and more focused regulatory learning on the specific field and, thus, enable specialization from the involved regulators[95]. This approach is expected to allow the provision of more precise guidance from regulators to organizations, thereby facilitating the efficient use of resources and expertise, generating empirical and measurable insights that inform clearer guidance and future regulatory adjustments[96].

However, unlike these sector-specific initiatives, this paper proposes to go one step further and establish requirement-specific sandboxes, with a particular focus on human oversight[97]. While sectoral sandboxes concentrate regulatory learning within a given industry or domain, a human-oversight sandbox would cut horizontally across sectors, targeting one of the most ambiguous and operationally challenging legal requirements contained in both the AI Act, the GDPR and other EU sectorial digital laws as applicable. Such a model would retain

88. Due to the fact that, in principle, national law may not derogate EU Law except where EU Law expressly allows for this, pursuant to the fundamental principle of primacy of EU law. For this reason, EU secondary law provisions on regulatory sandboxes such as those in the AI Act and Cyber Resilience Act are aimed at providing this flexibility, by allowing certain obligations to be derogated in the context of sandboxes, or to prevent the application of pecuniary fines by competent authorities. Bagni–Seferi 2025, pp. 29-43.

89. Commission Nationale de l'Informatique et des Libertés 2025, pp. 3-5.

90. Court of Justice, *Land Hessen/SCHUFA Holding AG*, case C-634/21.

91. See n. 18, *supra*.

92. See Section 2.1, *supra*.

93. Commission Nationale de l'Informatique et des Libertés 2025, p. 5.

94. For an overview of existing regulatory sandbox schemes in the EU and UK: Cirone 2025, pp. 263-266.

95. Lanamäki et al. 2025.

96. *Ibidem*.

97. A similar possibility has recently been suggested by Lanamäki et al. 2025 p. 41.

the benefits of specialization observed in national experiences – namely, more focused regulatory learning, concentration of expertise, and more precise guidance – but would channel them towards clarifying a single, cross-cutting obligation.

The advantage of establishing this type of sandboxes is intuitively apparent. By creating a safe space for regulators and AI operators to jointly probe how human oversight legal requirements may successfully be translated in practice, such sandboxes would primarily limit the elements to be tested in the sandbox to those related to Article 14 AI Act, hence facilitating targeted regulatory learning outcomes on how to implement effective and meaningful human oversight. The specialized, evidentiary feedback loop generated in a "human-oversight sandbox" could thus refine the interpretation of oversight obligations across legal instruments, ensuring that the legally mandated requirements of achieving – respectively – "effective" and "meaningful" oversight are translated into verifiable, evidence-based practices. In this respect, the operational advantage would lie in the possibility for the authorities operating the sandbox to specialize in human oversight, thus concentrating expertise, reducing the resources needed for fragmented or duplicative experimentation, and enabling faster and more efficient guidance to providers and deployers.

What is more, where relevant in light of the specific AI system tested within the sandbox, any overlapping or sector-specific human oversight provisions should likewise be incorporated in the sandbox perimeter. As noted above, such provisions can be found in other instruments pertaining to the EU digital acquis: for instance, an AI system aimed at allocating tasks to online platform workers would arguably fall into the scope of application of Article 14 AI Act (as a high-risk AI system under Annex III, point 4), Article 22 GDPR (as it processes workers' personal data to function), and Article 10 PWD, thereby triggering the human oversight requirements mandated by each of the three laws.

Despite the potential benefits, a first possible criticism is that the establishment of this kind of specialized sandboxes finds no justification in the relevant provisions of the AI Act. However, a closer look at the relevant provisions and recitals of the AI Act supports the view that specialized sandboxes for human oversight would not only be consistent with, but also directly advance, the objectives of the sandbox regime established under the law.

Not only no relevant provision of the AI Act on regulatory sandboxes appears to prevent the establishment of regulatory sandboxes that are "specialized" in one or more legal requirements, but several elements seem instead to leave open this possibility, and even to encourage it. For instance, in the context of the implementing acts which the Commission is mandated to adopt in order to complete the relevant framework for sandboxes, Article 58(2)(c) explicitly requires that AI regulatory sandboxes support, to the best extent possible, flexibility for national competent authorities to establish and operate their AI regulatory sandboxes. In this context, the concept of "flexibility" should be understood as encompassing the possibility of setting up both "general-purpose" sandboxes, covering the full set of legal requirements applicable to high-risk AI systems, and specialized ones focusing on particular obligations such as human oversight.

Even more importantly, Article 58(2)(i) mandates that sandboxes "facilitate the development of tools and infrastructure for testing, benchmarking, assessing and explaining dimensions of AI systems relevant for regulatory learning, such as accuracy, robustness and cybersecurity, as well as measures to mitigate risks to fundamental rights and society at large". Despite not being expressly cited in the exemplificatory list, human oversight is undoubtedly a key area where regulatory learning to the benefit of both regulators and providers can take place, since – as discussed previously – research and academic discourse show that effective oversight mechanisms remain elusive and highly contested in practice, requiring iterative, evidence-based experimentation to be meaningfully implemented. This is further compounded by Recital 139, which underlines that "participation in the AI regulatory sandbox should focus on issues that raise legal uncertainty for providers and prospective providers to innovate, experiment with AI in the Union and contribute to evidence-based regulatory learning". Again, as discussed in previous Sections of this paper, human oversight is an area where legal uncertainty is widespread – both due to practical issues and overlapping legal requirements stemming from EU law – and where, as a consequence, regulatory learning is much needed with a view to achieve legal certainty.

5.1. Governance and institutional design

However, in order to successfully meet their objectives, it is of paramount importance that the governance aspects are carefully considered, especially in light of the intersecting legal requirements around human oversight in EU law. The institutional setup of human oversight-specific sandboxes should thus foresee joint governance and responsibility between the competent AI market surveillance authority and the national DPA – which is always competent for human oversight requirements stemming from the GDPR and the Platform Work Directive[98] – with the possible involvement of further stakeholders, where relevant. In this latter respect, the role of competent standardization bodies that have been mandated by the EU Commission under Article 40 AI Act, to draw up technical standards which also encompass human oversight must also be emphasized. Within the New Legislative Framework, on which the AI Act is explicitly modelled[99], technical standards provide the concrete technical specifications through which essential legal requirements can be operationalized and verified, offering both state-of-the-art solutions and a presumption of conformity. This fundamental role of standards is all the more important in the field of human oversight, as they are expected to provide at least some degree of much needed operational guidance. Thanks to sandboxes, lessons learned from experimentation can directly feed into the standardization process, ensuring that emerging standards on open issues such as interpretability, robustness, and – crucially – human oversight, reflect real-world operational challenges. In this respect, the early involvement of standardization bodies in sandbox governance can help to align experimental practices with the forthcoming body of harmonized standards, enabling from the outset a feedback loop between regulatory learning and technical standard-setting[100]. This approach is supported by Article 58(2)(f) AI Act, which requires that implementing acts ensure the involvement of, *inter alia*, standardization organizations in the operation of AI regulatory sandboxes, thereby confirming their central role in this co-regulatory ecosystem.

5.2. Selection criteria and operational focus

Additionally, given the strict correlation between human oversight obligations applicable to providers under Article 14 and those incumbent upon deployers under Article 26[101], as discussed in Section 2.2 above, the possibility for providers to submit applications in partnerships with deployers, foreseen by Article 58(2)(b), should be especially emphasized and operationalized in the context of human oversight-specialized regulatory sandboxes. Bringing deployers into the sandbox environment adds considerable value, as they bear ultimate responsibility in ensuring that oversight mechanisms designed by providers are effectively translated into practice when AI systems are put to use. In line with this approach, Article 26 explicitly requires deployers to use high-risk AI systems in line with the provider's instructions for use, which – crucially – shall include human oversight measures[102], and to assign human oversight to natural persons with adequate competence, training, authority, and support. This structural dependency between the design obligations of the provider and the operational duties of the deployer can be tested to its full extent when both actors jointly participate in the sandbox. A cooperative approach between providers and deployers would thus not only facilitate a more realistic assessment of how human oversight operates in practice but also generate evidence-based insights to inform more coherent and actionable guidance for both categories of AI operators.

98. In particular, Art. 24(1) of the Directive states that: "The supervisory authority or authorities responsible for monitoring the application of Regulation (EU) 2016/679 shall also be responsible for monitoring and enforcing the application of Articles 7 to 11 of this Directive as far as data-protection matters are concerned […]".

99. For a thorough analysis, see: Veale–Zuiderveen Borgesius 2021.

100. Bagni–Seferi 2025.

101. Deployers may also be required to implement further oversight measures as data controllers or as digital labour platforms, where applicable, under, respectively, Article 22 GDPR and Article 10 of the Platform Work Directive, thus creating further complexity.

102. Art. 13(3)(c) AI Act.

To fully optimize the advantages of oversight-specific sandboxes, especially to foster legal certainty and regulatory learning, it is further suggested that competent authorities should operate strategically when considering which submissions to accept, by giving precedence to those applications that raise the most significant operative and/or interpretative challenges for the implementation of human oversight obligations. This approach would closely align with AI regulatory sandboxes' rationale and objectives: by prioritizing cases where the interpretation and/or operationalization of human oversight obligations are most uncertain, authorities would maximize the sandbox's added value as a vehicle for regulatory learning and evidence-based guidance for both competent authorities and AI operators. The European Commission's implementing acts on AI regulatory sandboxes, foreseen by Article 58 AI Act, should thus include the presence of legal barriers and interpretative challenges as an eligibility and selection criterion for prioritizing participation.

In practice, a human oversight-specific sandbox would operate not unlike a "traditional" sandbox, that is, through phased cycles (typically, application phase, testing and experimentation, and exit), for a limited time and with the issuance of a final exit report, to be leveraged by participants to demonstrate compliance with the relevant obligations and to be accessed by the Commission and other competent authorities[103]. These reports would moreover produce reusable deliverables and should hence preferably be made public to enhance regulatory learning outcomes, as foreseen by Article 57(8) AI Act: in case where the individual sandbox application is successful, they may contain practical insights and guidelines for the application of human oversight requirements, such as organizational playbooks to carry out oversight activities successfully, training materials for overseers, best practices for drafting and implementing providers' instructions for use, standardized documentation formats and potential suggestions for amendments to current laws and technical standards could emerge from sandbox testing. At a minimum, insights into best practices and lessons learned from operating the sandbox should be included in the yearly reports to be sent by competent authorities to the AI Office under Article 57(16) AI Act. Ideally, to further favour their dissemination and uptake across the Union, such insights should also be integrated into the AI literacy initiatives coordinated by the AI Office, thereby ensuring that the lessons learned translate into broader capacity-building for both regulators and AI operators.

On the operational side, the specific sandbox plan to be agreed *ex ante* between competent authorities and the applicant should specify: (i) the chosen model of oversight (e.g., human-in-the-loop/on-the-loop/in-command, or a combination thereof) together with its technical measures and safeguards (e.g., granular logging to ensure accountability, kill switches, threshold alerts for bias detection, minimum interpretability requirements); (ii) organizational procedures for the allocation of oversight roles, to be included by the provider in the instructions for the deployer, recommended training requirements for human overseers (also in light of AI literacy requirements under Article 4 AI Act), maximum review times, traceability of overrides or deviations; (iii) metrics to measure oversight effectiveness, including error detection rates, justified overrides, timeliness of interventions, consistency of decision-making, levels of unlawful bias mitigated, and indicators of automation reliance; and (iv) in light of the overlapping GDPR requirements, a Data Protection Impact Assessment with a focus on data protection-mandated human oversight requirements under Article 22[104].

If carefully planned, also in light of the aforementioned policy recommendations, specialized sandboxes for human oversight could play a decisive role in addressing the gap between the EU legislator's aspirations on human oversight and the real-world shortcomings thereof. Their introduction would thus represent a crucial step towards transforming human oversight from a largely rhetorical safeguard into a verifiable practice.

The following conclusions will summarize the main findings of this paper and point to possible avenues for further research and policy development.

103. See, respectively, Art. 57 paras. 7 and 8.

104. A clear example of such processing activities would be the generation and retention of logs related to the human overseer's activities, as also mandated by Arts. 12 and 26 of the AI Act.

6. Conclusions and path forward

Drawing from current literature and research, this paper has argued that while human oversight has become one of the EU legislator's leading safeguards against algorithmic harms, its practical effectiveness is far from guaranteed and, if not carefully implemented, may even worsen already-present issues. Cognitive, organizational, and structural shortcomings, together with regulatory overlap, undermine the very safeguard that the law places so much trust in. In this respect, regulatory sandboxes – as foreseen in the AI Act – offer a promising venue to transform human oversight from a rhetorical safeguard into a verifiable, evidence-based practice.

The analysis carried out in Section 3 has shown that while human oversight remains one of the central safeguards in EU regulation of automated decision-making, its practical implementation is not without its shortcomings. Cognitive limitations, organizational pressures, and systemic biases all compound to render oversight often more symbolic than effective, thus undermining the very objective of the GDPR and AI Act to protect fundamental rights when deploying automated decision-making. It is against this problematic scenario that specialized AI regulatory sandboxes for human oversight emerge as a particularly promising avenue to reap the full benefits that are promised by human oversight measures. This kind of tailored sandboxes would foster "specialized" regulatory learning on human oversight, generate empirical evidence to inform coherent, EU-wide guidance and technical standardization in this sector, both to shape current interpretation and understanding of legal measures on oversight, and to consider possible future adjustments to the legal framework.

In this respect, the forthcoming European Commission's implementing acts under Article 58 AI Act, offer a unique opportunity to embed this specialization into the sandbox framework. The Commission should ensure that the implementing rules explicitly allow sandboxes to focus on specific legal requirements such as human oversight and promote the systematic dissemination and mutual recognition of sandbox results across Member States.

Specialized sandboxes tailored for human oversight could help regulators and AI operators experiment with concrete models, metrics, benchmarks, and practices to actually meet legally-mandated requirements for human oversight, while also enabling regulatory learning from involved authorities. In this latter respect, by concentrating and stimulating specialized expertise in competent authorities, reducing fragmented experimentation, and directly feeding lessons learned into forthcoming technical standards, common specification and AI literacy initiatives, these sandboxes could accelerate the correct operationalization of horizontal human oversight obligations under Articles 14 and 26 AI Act, Article 22 GDPR, as well as other sectorial and overlapping human oversight obligations, notably Article 10 of the Platform Work Directive, including their overlapping application.

At a broader policy level, sandboxes also represent an important opportunity for the EU to reaffirm its fading global leadership in digital regulation at a moment when that leadership is being questioned. As underlined in the Draghi Report, excessive and overlapping regulation in the digital sphere risks eroding the Union's competitiveness[105]; yet, the same report singles out sandboxes as innovation-friendly mechanisms that can enable continuous assessment of regulatory hindrances[106]. If the EU manages to effectively operationalize human oversight through specialized sandboxes, it could decisively influence other jurisdictions currently adopting (and, possibly, struggling with the operationalization of) similar requirements, thus regaining its lost momentum in setting global standards for digital regulation.

Member States, and even regional or local authorities, have a unique opportunity to build sandboxes dedicated to human oversight, possibly linked to sector-specific experimentation frameworks (e.g., those already existing in the field of FinTech, healthcare, automotive, etc.). This could create an EU network of oversight sandboxes that leverages diversity of contexts while ensuring coherence through coordination by the AI Office under Article 57 AI Act.

105. Draghi 2025, p. 79.

106. *Ivi*, p. 84.

This paper aims to stimulate a debate on whether such an approach is desirable and, if so, how it could be structured and governed, also offering some recommendations and policy proposals. Much will depend on the political will and resources that the EU and its Member States devote to these instruments. In any case, this path seems preferable than the current drift towards deregulation in EU digital policy which has taken place in the last year, amidst the Draghi Report and geopolitical tensions with the United States[107], as it provides a venue for the EU to maintain and even advance its regulatory leadership on digital regulation.

Further research should provide a closer look on how experimentation within regulatory sandboxes can best support the adoption and testing of human oversight requirements, in particular by operationalizing Article 14 AI Act, and other oversight obligations through forthcoming technical standards or common specifications, and by identifying measurable benchmarks and metrics for oversight effectiveness across different sectors[108].

References

A. Adinolfi (2022), *Processi decisionali automatizzati e diritto antidiscriminatorio dell'Unione europea*, in A. Adinolfi, A. Simoncini (a cura di), "Protezione dei dati personali e nuove tecnologie – Ricerca interdisciplinare sulle tecniche di profilazione e sulle loro conseguenze giuridiche", Edizioni Scientifiche Italiane, 2022

Article 29 Working Party (2017), *Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679 (Wp251rev.01)*, 2017

F. Bagni (2023), *The Regulatory Sandbox and the Cybersecurity Challenge: From the Artificial Intelligence Act to the Cyber Resilience Act*, in "Rivista Italiana di Informatica e Diritto", 2023, n. 2

F. Bagni, F. Seferi (Eds.) (2025), *White Paper on Regulatory Sandboxes for AI and Cybersecurity*, CINI's Cybersecurity National Lab, 2025

S. Barros Vale, G. Zanfir-Fortuna (2022), *FPF Report: Automated Decision-Making Under the GDPR – A Comprehensive Case-Law Analysis*, in Future of Privacy Forum, 2022

K. Brennan-Marquez, D. Susser, K. Levy (2019), *Strange Loops: Apparent versus Actual Human Involvement in Automated Decision-Making*, in SSRN Scholarly Paper, 2019

S. Cansu, J. Farrar (2021), *Managed by Bots Report*, in "Worker Info Exchange", 2021

E. Cirone (2025), *Gli spazi di sperimentazione normativa nell'Unione europea: regolamentare l'innovazione tra principi e prassi applicative*, in "Rivista Italiana di Informatica e Diritto", vol. 7, 2025, n. 1

D. Collingridge (1980), *The Social Control of Technology*, Frances Pinter Publisher Ltd., 1980

Commission Nationale de l'Informatique et des Libertés (2025), *Artificial intelligence and public services: the CNIL publishes the results of its "sandbox"*, in cnil.fr, 18 April 2025

Council of Europe (2024), *Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law,* 2024

J. Dirutigliano, D. Baldini (2025), *The Right to Explanation: Legal Challenges and the Future of Fairness in Automated Decision-Making*, in "Journal of AI Law and Regulation", vol. 2, 2025, n.2

107. Politico.eu 2025.

108. An example of metrics concerning human oversight effectiveness in reducing (or indeed worsening) algorithmic discrimination has been devised and deployed by Gaudeul et al. 2025. Other metrics could be developed and applied in other contexts, for example in order to measure human oversight effectiveness in reducing safety incidents for highly autonomous systems, and so on, to be tested in a specialized regulatory sandbox.

L. Downes (2009), *The Laws of Disruption: Harnessing the New Forces That Govern Life and Business in the Digital Age*, Basic Books, 2009

M. Draghi (2025), *The Future of European Competitiveness – In-depth analysis and recommendations (Part B)*, 2025

L. Enqvist (2023), *Human Oversight, in the EU Artificial Intelligence Act: What, When and by Whom?*, in "Law, Innovation and Technology", vol. 15, 2023, n. 2

European Commission (2023), *'Better regulation' toolbox*, July 2023 edition

M. Fink (2025), *Human Oversight under Article 14 of the EU AI Act*, in SSRN Scholarly Paper, 2025

A. Gaudeul et al. (2025), The Impact of Human-AI Interaction on Discrimination, Publications Office of the European Union, 2025

B. Green (2022), *The Flaws of Policies Requiring Human Oversight of Government Algorithms*, in "Computer Law & Security Review", vol. 45, 2022

B. Green, A. Kak (2021), *The False Comfort of Human Oversight as an Antidote to A.I. Harm*, in SSRN Scholarly Paper, 2021

High-Level Expert Group on Artificial Intelligence (2019), *Ethics Guidelines for Trustworthy AI*, 2019

H.M. Holtz, J. Ledendal (2026), *AI Data Governance – Overlaps Between the AI Act and the GDPR*, in "Law, Innovation and Technology", 2026 (forthcoming)

A. Lanamäki et al. (2025), *What to Expect from the Upcoming EU AI Act Sandboxes: Panel Report*, in "Digital Society" vol. 4, 2025, n. 2

G. Lazcoz, P. De Hert, (2022), *Humans in the GDPR and AIA Governance of Automated and Algorithmic Systems. Essential Pre-Requisites against Abdicating Responsibilities,* in SSRN Scholarly Paper, 2022

E. Longo (2021), *Time and Law in the post-COVID-19 Era: the usefulness of Experimental Law*, in "Law and Method", Special Issue: Experimental Legislation in Times of Crisis, edited by S. Ranchordàs, B. van Klink, 2021

T. Madiega, A.L. Van De Pol (2022), *Artificial intelligence act and regulatory sandboxes*, EPRS – European Parliamentary Research Service, 2022

T. Moraes (2023), *Regulatory Sandboxes as Tools for Ethical and Responsible Innovation of Artificial Intelligence and Their Synergies with Responsive Regulation*, in SSRN Scholarly Paper, 2023

OECD (2024), *Recommendation of the Council on Artificial Intelligence*, OECD Legal Instruments, 2024

OECD (2023), *Regulatory Sandboxes in Artificial Intelligence*, OECD Digital Economy Papers, 2023

A. Panezi (2024), *Requirements of High-Risk AI Systems: AI Act. Article 14. Human Oversight*, in SSRN Scholarly Paper, 2024

Politico.eu (2025), *Trump threatens 'substantial' new tariffs against countries with 'discriminatory' digital rules*, 2025

S. Ranchordas, V. Vinci (2024), *Regulatory Sandboxes and Innovation-Friendly Regulation: Between Collaboration and Capture*, in "Italian Journal of Public Law", vol. 1, 2024

D.J. Solove, H. Matsumi (2024), *AI, Algorithms, and Awful Humans*, in "Fordham Law Review", vol. 92, 2024, n. 5

L. Tosoni (2021), *The right to object to automated individual decisions: resolving the ambiguity of Article 22(1) of the General Data Protection Regulation*, in "International Data Privacy Law" vol. 11, 2021, n. 2

UNESCO (2021), *Recommendation on the Ethics of Artificial Intelligence*, UNESDOC Digital Library, 2021

United Nations (2024), *Resolution on AI*, United Nations General Assembly Resolutions, 2024

M. Veale, F. Zuiderveen Borgesius (2021), *Demystifying the Draft EU Artificial Intelligence Act – Analysing the Good, the Bad, and the Unclear Elements of the Proposed Approach*, in "Computer Law Review International", vol. 22, 2021, n. 4

S. Wachter (2022), *The Theory of Artificial Immutability: Protecting Algorithmic Groups under Anti-Discrimination Law*, in "Tulane Law Review", vol. 97, 2022, n. 2