Research Ethics Governance with Responsible AI Sandboxes

Michael Gille, Marina Tropmann-Frick

Hamburg University of Applied Sciences, Germany

DOI 10.3217/978-3-99161-062-5-010, CC BY 4.0

Abstract. University research ethics committees (REC) face challenges in overseeing artificial intelligence (AI) research. Historically rooted in biomedical and social science paradigms, REC were not designed to evaluate the epistemic, temporal, and normative complexities of AI and machine learning research. The EU's AI Act exacerbates this tension by exempting academic research from its scope while at the same time promoting the application of ethics guidelines, thereby creating a zone of normative ambiguity. This paper critically examines the resulting governance vacuum. We argue that conventional ethical review processes are inadequate in many cases for reasons inherent in AI research, which is often iterative and interdisciplinary, characterized by shifting goals and emerging risks, as well as because of the normative and socio-technical co-construction of AI technology development. We propose the *Responsible Artificial Intelligence Sandbox* as a model for research ethics governance. It reframes the role of REC from static evaluators to co-constructors of ethical oversight within experimental research environments. Drawing on insights from regulatory sandboxes in EU law and national contexts, this conceptual model enables dynamic, participatory, and reflexive engagement with ethics throughout the research lifecycle. Two main contributions are made: we diagnose structural misalignments of existing research ethics infrastructure and conceptualize responsible AI sandboxes as an institutional and methodological innovation that aligns ethical governance with the nature of research on and with AI.

1 Introduction

University-based Research Ethics Committees (REC), Institutional Review Boards (IRB), and Ethics Review Committees (ERC), in the following collectively referred to as REC, are set up to ascertain ethically responsible research oversight. Rooted historically in biomedical, behavioural, and social research paradigms, these bodies were institutionalized to protect human subjects and uphold normative standards of scientific integrity (Shamoo and Resnik, 2009). However, the rapid emergence of data processing, algorithmic, machine/deep learning (in the narrower sense) and artificial intelligence (AI) research (in the wider sense), hereafter jointly referred to as AI research, have exposed

profound limitations in established research ethics governance models (Stahl et al., 2025; Hadley et al., 2025; Bouhouita-Guermech et al., 2023; Petermann et al., 2022; González-Esteban and Calvo, 2021; Ferretti et al., 2021). As computational systems become both the objects and instruments of inquiry, REC increasingly face tasks they have initially neither conceptually nor procedurally been designed to address. At the same time, the European legislator takes a cautious approach to AI research and grants university research far-reaching freedom. The EU's Artificial Intelligence Act (AI Act), introducing a risk-based regulatory framework for AI development and deployment, explicitly exempts academic AI research from its direct applicability (AI Act, Art. 2(6), rec. 25). This exemption produces a structurally intended zone of governance ambiguity. Mandated to maintain the high standards of biomedical ethics reviews (Brenneis et al., 2024), universities thus find themselves in a paradoxical position: they are encouraged to advance high-risk AI research under conditions of ethical autonomy, while lacking institutional mechanisms and resources tailored to the novel and recursive dilemmas this research entails.

AI research often involves open-ended exploration, emergent objectives, and shifting definitions of key principles such as fairness, trustworthiness, explainability, and human-centeredness (Díaz-Rodríguez et al., 2023). In such volatile research settings, ethical and legal oversight cannot be a static, one-time assessment. Yet traditional review processes struggle to keep pace, often reduced to bureaucratic hurdles by technical researchers, or overwhelmed by the sheer complexity of innovation (Brenneis and Burden, 2024). The governance landscape is increasingly saturated with normative frameworks, ranging from 'trustworthy AI' to data protection principles and fundamental rights mandates (Jobim et al, 2019), but these frameworks often operate in silos, lack enforceability, or offer only abstract guidance (Resseguier, 2024; González-Esteban and Calvo, 2024). What is absent is a tailored mode of ethics governance that is dynamic, participatory, and capable of engaging with uncertainty that research 'on' and 'with' AI brings about. This paper proposes the concept of the *Responsible AI Sandbox* as a model of integrative research ethics governance. Sandboxes, as spaces for regulatory and technical experimentation, offer an architectural shift: they enable REC not only to assess, but to enable the *co-construction* of ethical and governance practices within the experimental settings of AI innovation itself, while at the same time reducing the risk of evaluation gaps.

This paper makes two contributions to ongoing debates in science and technology studies (STS), research governance, and AI research ethics, providing specific impetus for a normative reflection and governance of AI research (research that develops and/or uses AI): *First,* it critically diagnoses the dilemmatic structural inadequacies of traditional ethics-approval mechanisms when confronted with the epistemic, temporal, and normative complexities of AI research. In this regard, special attention is paid to the co-constructive character of responsible AI research, in which the very definitions of risk,

fairness, trustworthiness, and accountability are shaped within the research process itself, rather than being fully specifiable ex ante. By stressing this entanglement of ethics and epistemics, the paper contributes to STS and RRI scholarship on the co-production of algorithmic knowledge and normativity, while offering a fresh institutional diagnosis of selected pressures REC currently face. *Second,* the paper introduces and conceptualizes the *Responsible Artificial Intelligence Sandbox* (RAIS) as an integrative governance model that creates not just a space to engage with ethical norms, but enables dynamic, participatory, and reflexive forms of oversight within university AI research settings. At the same time, this sandbox model is positioned as a response to the governance vacuum created by the EU AI Act's exemption of research activities, and as a concrete mechanism for enabling regulatory learning through an open and normatively reflective research practice within academic institutions. Framed as both a methodological and institutional innovation, this sandboxing approach allows universities to experiment with ethical and legal frameworks together with technological development, thus transforming them into laboratories of governance modalities themselves.

This paper is guided by a set of research questions articulating a broader inquiry into how universities can adapt their governance infrastructures to align with the epistemic and normative demands of responsible AI research, while remaining grounded in principles of academic freedom and anticipatory accountability. The following questions structure the inquiry:

- What *structural* limitations and *epistemic* mismatches do REC face when tasked with the oversight of AI research?
- How should a 'Responsible AI Sandbox' be conceptually designed to meet key ethical and regulatory imperatives such as responsibility, fairness, trustworthiness, and risk mitigation and what insights can, in this regard, be drawn from regulatory sandbox approaches in EU law and national jurisdictions through analogical reasoning and governance transfer?

This paper employs a conceptual research methodology grounded in interpretive STS and regulatory studies, synthesising insights from policy documents, scholarly literature including comparative case analyses of regulatory sandboxes. It uses abductive reasoning to explore how institutional design elements can be transposed into intra-university governance frameworks for responsible AI. The approach highlights normative and epistemic dimensions of experimentation, focusing on co-construction, reflexivity, and anticipatory governance. Through critical analysis, analogical reasoning and institutional comparison, the paper seeks to translate and adapt principles of experimental regulation to the university context, with the aim of developing a conceptual framework for responsible AI sandboxes.

This paper is organized as follows. Section 2 identifies relevant strands of literature on the ethical governance of AI research. This is followed by methodological considerations (section 3). The subsequent section 4 lays out selected structural pitfalls in the ethical governance of AI research before the paper then turns to the responsible AI sandbox (section 5). We conclude with an outlook in section 6.

2 Related Work

This paper´s point of departure are approaches and reflections on AI research ethics governance. Owing partly to AI research leading to critical scrutiny and criticism of the 'traditional' research ethics governance set-up in universities, there have been calls for reforms (Masso et al., 2025; Gonzáles-Esteban and Calvo, 2022; Petermann et al., 2022). Increasing attention is being paid to the shifting role of REC in AI-related research, regarding both research on and with AI (Esmaili et al., 2025; Brenneis/Burden, 2024; ZEVEDI, 2023). For purposes of this paper´s integrative approach, two related strands of discussion stand out and are put into focus: There are calls for an interdisciplinary adaptation and expansion of existing REC by integrating computer and data science into existing REC bodies, including experts from additional disciplines and training (Stahl et al., 2025; Brenneis et al., 2024) and introducing new principles and guidelines (Bouhouita-Guermech et al., 2023; Hagendorff, 2020). Apart from notions of such 'Super-REC', the additional creation of specialized sub-committees is advocated (Esmaili et al., 2025). Deviating from this idea of `traditional´ REC assuming additional tasks there is also a strand of literature that puts forth the idea additional separate and specialized REC, dubbed, e.g., as Algorithmic Research Ethics Committee/Board (ARB) or AI Research Committee (AIRC) (Hadley et al., 2024; Jordan, 2019).and including flexible approaches such as ETHNA (González-Esteban and Calvo, 2022).

Our analysis further builds on scholarship that explores the co-constructivist nature of techlaw, i.e. the idea that both, hard and soft law norms evolve together with technological development in mutual dependency (Jones, 2018; Crootof and Ard, 2021). This approach emphasizes the formative role of norms within innovation environments and offers a theoretical grounding for designing regulatory sandboxes as sites of iterative governance rather than merely reactive or pre-emptive control. This research also allows for an inclusion of the EU AI Act´s risk-based approach into the AI research governance (Resseguier and Ufert, 2024; Wernicke and Meding, 2025).

Since REC are questioned as the optimal structure for AI research ethics oversight (Stahl et al., 2025), we consider the growing body of scholarship that examines the emergence and operationalization of regulatory sandboxes as instruments for normative experimentation and 'moral imagination' (Undheim et al., 2022): This includes research on their role in the EU's AI Act, where regulatory sandboxes are proposed as controlled

environments for innovation under public supervision, focusing on their potential to balance innovation with regulatory oversight (Plato-Shinar and Godwin, 2025; Undheim et al., 2022; Ranchordas, 2021). While regulatory sandboxes have been primarily understood as tools for innovation governance at the state or market level (Gumbo and Chude-Okonkwo, 2025), we build on ethical approaches to regulatory sandbox conceptualizations (Francis, 2025) and propose an integration of responsible AI frameworks (Göllner et al., 2024a). Our approach can therefore be categorized as 'intra method', as it primarily deals with the responsible design of technology (Reijers et al. 2018).

3 Methodology

This article adopts a conceptual and analytical methodology, combining theoretical inquiry with normative analysis. The study draws on interdisciplinary frameworks, primarily from STS, responsible AI research and legal theory, to examine the underlying logic and implications of an AI research ethics sandbox based on responsible AI considerations. Through this approach, the paper aims to clarify key concepts, identify underlying assumptions, and develop a structured argument based on existing literature and normative reasoning. This approach is underpinned by the notion of reflexive governance (Voss & Kemp, 2006; Feindt et al., 2018), which emphasizes iterative policy development, stakeholder deliberation, and the institutionalization of uncertainty. Such a perspective is especially pertinent to the AI research domain, where the consequences of methodological and technical decisions are often opaque, distributed between research disciplines and stakeholders, and temporally deferred. To this methodological end, the paper employs analogical reasoning and governance learning (Stone, 2012; Sabel and Zeitlin, 2012; Rangoni, 2022) to identify structural and functional parallels between regulatory sandboxes in innovation policy and the exigencies of ethical review mechanisms in AI research governance. The approach involves the mapping of problem similarities (e.g., uncertainty, novelty, rapid change), institutional roles (e.g., regulatory gatekeepers, facilitators, enablers), and process characteristics (e.g., iterative evaluation, stakeholder feedback loops). This 'cross-pollination' with sandbox logic opens a space for reflexive governance (Voss & Kemp, 2006), wherein ethical oversight becomes an iterative and participatory process rather than a fixed ex ante assessment. By integrating these methodologies, we purpose a conceptual basis for rethinking how academic research governance can responsibly adapt to emerging challenges such as AI or algorithmic experimentation as well as data-intensive applications, transferring regulatory sandbox logic to AI research ethics oversight.

4 Pitfalls in the Ethical Governance of AI Research

4.1 The Role of Research Ethics Committees in Research on and with AI

REC are entrusted with safeguarding ethical standards and balancing academic freedom with societal responsibility. Their role has become increasingly complex in the context of AI research, where both the pace and epistemic configuration of research challenge the assumptions underpinning conventional review processes of REC (Gonzáles-Esteban and Calvo, 2022; Esmaili et al., 2025). The normative principle of *rule-bound academic freedom* demands that universities, while enjoying institutional autonomy, demonstrate responsibility in overseeing potentially harmful research (Wernick and Meding, 2025). REC embody this responsibility, yet their practices remain rooted in paradigms often not doing justice to the technical and systemic features of AI. AI research often implicates diffuse, systemic, unpredictable harms, whether to privacy, fairness, or fundamental rights, issues that exceed the scope of traditional ethics assessment (Jobin et al., 2019; Mittelstadt, 2022).

The governance challenges differ markedly between *research on AI*, i.e. research which targets AI as its object and usually involves algorithmic development, and *research with AI*, where, often proprietary, AI algorithms serve as tools for inquiry in other domains and disciplines (Stahl et al., 2025). The former entails direct engagement with the design, testing, and evaluation of AI systems, while the latter embeds AI within disciplinary contexts where its limitations and embedded values may remain obscured. REC are increasingly tasked with reviewing both types, often without tailored methodologies or a shared vocabulary of risk.

4.2 Structural Critique: Why the Traditional Ethics Review Procedure Falls Short in AI Research

Structured around linear application and approval procedures, the traditional ethics review model often is unsuitable for AI research. REC were designed to assess clearly scoped studies with stable methods and foreseeable risks. AI research, by contrast, often unfolds within iterative, interdisciplinary, and exploratory projects whose normative and epistemic contours emerge during the research process itself (Stahl et al., 2025). In many AI and big data projects, key ethical dimensions, especially fairness, trustworthiness, human-centeredness, bias mitigation, and explainability, are not predefined checkboxes but outcomes of ongoing technical, conceptual, and empirical work (Gonzáles-Esteban and Calvo, 2022). The same applies to fundamental rights implications, which are frequently discovered or clarified only through experimentation (Bouhouita-Guermech et al., 2023; Díaz-Rodríguez et al., 2023). This `epistemic fluidity´ produces a structural tension: REC are expected to conduct anticipatory assessments of research that involves algorithmic and non-algorithmic rules yet to be fully articulated. As a result, ethics

oversight risks devolving into a formalistic exercise, imposing rigid scrutiny on processes that require ongoing adaptive reflection, resulting in review gaps (Reijers et al., 2018; Zimmer, 2018). This structural dilemma is amplified by the AI Act's explicit exemption of academic research from its binding regulatory scope, leaving REC as de facto ethical gatekeepers. However, they are poorly equipped for this role without appropriate institutional tools or temporal flexibility. The principle of precaution itself is put at risk: premature or shallow review may inadvertently sanction a normative vacuum, with ethical reflection sidelined until after deployment or publication.

These challenges are not merely theoretical and touch upon almost all disciplines and modes of AI use (Brenneis and Burden, 2025; Masso et al., 2025). Issues arise, e.g., from proprietary software, opaque model behaviors, and the re-identification potential of anonymized datasets (Esmaili et al., 2025). Across diverse cases one pattern recurs: the intertwining of data and algorithmics generates ethical risks that extend far beyond data protection. These include proprietary model opacity, normatively charged classification decisions, and epistemic displacement, where complex social judgments are outsourced to probabilistic systems. Such risks cannot be adequately anticipated in advance because they emerge from the interplay of technical architecture, data provenance, and research context.

Moreover, when AI functions as an epistemic infrastructure rather than as an object of inquiry, its normative implications risk becoming invisible. This is particularly problematic in applied fields, where AI tools are operationalized without critical reflection on their embedded assumptions. In such fast-moving research environments, requiring formal review for every AI application is impractical and potentially stifling. What is needed instead is an adaptive, recursive governance model, i.e. oversight that enables parallel ethical reflection.

4.3 Co-Constructing Responsibility

Debates about the governance of emerging technologies are often framed around a reactive normative paradigm: hard law and soft law perpetually scrambling to catch up with the speed of innovation (Calo, 2015). This 'normative lag' narrative, emblematic of what Jones (2018) refers to as 'technological exceptionalism', positions normative frameworks as outdated or inert in the face of rapid technological transformation. But this framing misrepresents the complex reciprocity between technical and normative systems. The governance of AI research cannot be meaningfully addressed through a simple linear model in which (ethical) norms react to technological change. While it is often said that rules 'lag behind' innovation, this narrative of reactive governance obscures a more fundamental dynamic: technology and regulation are mutually constitutive (Jones, 2018; Kaminski, 2023). Norms do not merely constrain or respond to technological development; they are co-produced with it, shaping what is built, how it is tested, and what is ultimately seen as acceptable or desirable. According to this

scholarship, technology can be characterized as a *socio-legal construction*, a view that waives the idea of technology as a neutral or autonomous force and instead emphasizes the normative architectures in which it is embedded from the start. Norms are to be seen as an infrastructural component of innovation, involved in everything from institutional design to the allocation of liability and legitimacy (Crootof and Ard, 2021). Seen through this lens, rules are not merely a set of exogenous constraints: One of the defining characteristics of AI research is its experimental, iterative, and exploratory nature. Many projects do not begin with fixed hypotheses, stable methodologies, or clearly anticipated outcomes (Esmaili et al., 2025). Instead, they involve speculative inquiry into algorithmic behaviour, emergent model properties, or complex data interactions. As such, key normative categories such as trustworthiness, fairness, transparency, bias mitigation, human-centeredness, are not static benchmarks to be checked off but are formed and refined through the research process itself.

The co-constructive nature of AI research governance is relevant for both, *research on* and *research with* AI. In research *on* AI, normative concerns are often an explicit part of the development process, as researchers examine issues such as algorithmic bias, model robustness, or the implications of scale and generalization. In contrast, research *with* AI uses AI systems as tools or infrastructures that support inquiry in other disciplines. Whether used in medical diagnostics, nursing, or historical text analysis, AI becomes a background instrument. Yet precisely in these settings, the embedded assumptions, data dependencies, and potential harms of AI systems can become invisible (Masso et al., 2025). Normative issues, ranging from bias and opacity to re-identification risks, may go unexamined or at least not understood sufficiently because AI is seen not as a subject of scrutiny, but as a technical utility.

This ‘rules-in-the-making’ logic creates a structural challenge for traditional ethical governance mechanisms. REC, while essential for safeguarding rights and ensuring accountability, are often tasked with prospectively assessing projects against normative standards that are not yet clear. This creates a temporal and epistemic mismatch: ethics oversight mechanisms are expected to anticipate and evaluate risks that are still unfolding and often unknowable in advance. This is not a matter of regulatory failure but of structural incompatibility: AI research generates its own norms as it proceeds, particularly in domains such as fairness and explainability, where solutions are context-sensitive and often co-produced in dialogue with technical, disciplinary, and societal inputs. REC, designed around anticipatory governance, find themselves at the limits of their institutional design: asked to adjudicate the ethical soundness of research trajectories whose parameters are still under construction. A co-constructive understanding of AI research governance thus requires moving beyond linear or top-down models of oversight.

Responsibility is not a pre-defined standard to be enforced ex ante; it is an evolving, situated, and distributed practice. Researchers, technologists, legal experts, ethicists, and communities must, depending on the research, participate in shaping what responsibility means in context, i.e. across different phases of research, applications, and institutional settings (Stilgoe et al., 2020; König et al., 2021). This reframing also calls for rethinking the institutional role of ethics governance bodies. Rather than serving primarily as gatekeepers issuing once-off approvals, REC might serve their role better if they also facilitate ongoing normative reflection, and, in so doing, support researchers in articulating, contesting, and refining the values that guide their work (Masso et al., 2025). This requires institutional learning mechanisms, interdisciplinary dialogue, and openness to the provisional nature of ethical judgments in complex, high-uncertainty domains like AI. In short, norms do not merely follow technology, nor can ethics be 'applied' to research like a seal of approval. Both are part of the architecture of innovation itself. In responsible AI research, norms and systems must be developed together, in an iterative and participatory manner that acknowledges their mutual dependencies (Göllner et al., 2024b). Only by recognizing and institutionalizing this co-constructive dynamic governance frameworks can be developed that are genuinely capable of meeting the ethical and epistemic challenges of AI research.

4.4 Risk-based Approach to AI Research Ethics Assessment

With its risk-based framework, the EU AI Act offers a useful, albeit indirect, point of reference for soft law approaches (Gille et al., 2024). Though the AI Act exempts academic research covering AI systems and AI models, including their output, from its formal scope (AI Act, Art. 2(6), rec. 25), its emphasis on risk classification, harm mitigation, and fundamental rights protection provides REC with an external source of normative orientation (Resseguier and Ufert, 2024; Wernicke and Meding, 2025). Complementary frameworks, such as the European Commission's High-Level Expert Group (HLEG) on AI and its 'Ethics guidelines on trustworthy AI' (HLEG, 2019; Göllner et al, 2024a), offer further guidance for the formulation of internal ethics policies and procedural criteria.

AI research invariably operates under conditions of uncertainty, where risks, technical, ethical, social, and systemic, are often diffuse and emergent. In this landscape, legal and ethical standards cannot be cleanly codified in advance but must evolve with and through technological practice. At the same time, risk is not an anomaly to be eliminated but a constitutive feature of the policy choice underpinning risk-oriented regulation of digital innovation, a 'risk baggage' (Kaminski, 2023). This constitutive aspect demands a structural response: risk mitigation must be integrated into research processes, not appended after the fact, for the research is in many cases tested in real-world situations and is often likely to end up in applications in market environments. This aspect goes even further, namely in ethics-by-design methodologies ('EbD-AI'), i.e. the

comprehensive and systematic inclusion of normative-ethical considerations in the design and development of AI (d'Aquin et al., 2018; Brey and Dainow, 2023). Such ethics-informed design and development considerations would also bring the research output into line with notions of fundamental rights impact assessment (FRIA) brought in by the EU AI Act (Mantelero, 2024).

5 The Responsible Artificial Intelligence Sandbox

5.1 Responsible Artificial Intelligence in the Research Ethics Review Process

Not all AI research projects require formal approval by REC. A categorical distinction is necessary to preserve both the integrity of ethics governance and the operational feasibility of review processes. A differentiated approach, based on technical and epistemic considerations, enables a more targeted allocation of ethical oversight and is tentatively delineated as follows:

1. Research activities that involve human subjects, process sensitive personal data, are security-related, or have foreseeable implications for individual rights and social outcomes, must remain subject to the standard REC procedure. This includes studies deploying AI in projects, e.g., involving biometric recognition, behavioural prediction, or automated decision-making with high-stakes consequences. The same applies also to any research that evaluates or calibrates AI systems using personally identifiable or vulnerable data.
2. Purely technical or foundational research on AI models that does not involve human participants, personal data, or application scenarios with immediate ethical relevance can be excluded from mandatory REC review. This research includes algorithm development, benchmarking on synthetic or anonymized datasets, formal model analysis, or optimization/refinement studies where no direct or indirect harm is plausible. Requiring ethics review for such abstract work would prevent unnecessary procedural overhead.
3. Between these two categories lies a ‘grey zone’ of interdisciplinary and application-oriented AI research that operates under conditions of epistemic uncertainty and dynamic normative standards. Such research, e.g., using pretrained models in new domains, combining social datasets with machine/deep learning techniques, or deploying AI in exploratory decision support scenarios, often does not initially meet the criteria for formal REC review, yet may develop ethical risks over time.

Responsible AI sandboxes can address the issues of the third category and allow researchers to conduct reflexive assessments dynamically adjusting review thresholds.

This way a sandbox supports sensitivity analyses across different models and application domains, enabling the systematic ethics evaluation.

5.2 Responsible Artificial Intelligence Sandbox

5.2.1 A sandbox for Responsible Artificial Intelligence Research and Innovation

The concept of the sandbox, borrowed from software engineering and regulatory experimentation, offers a productive frame to resolve challenges outlined in the previous sections. Where technological sandboxes isolate code from production environments to allow for safe experimentation, regulatory sandboxes extend this notion to legal and ethical governance, providing time-bound, supervized environments for controlled testing under conditional flexibility (Ranchordas, 2021; Longo and Bagni, 2025). The concept of the regulatory sandbox, increasingly common in technology regulation, has gained traction as a model for iterative, adaptive governance under uncertainty. However, to address the distinctive complexities of AI research within academic settings, a further evolution is required: the development of a Research and Innovation Sandbox for responsible AI research.

Regulatory sandboxes provide a practical instantiation of the co-constructive logic outlined in section 4. Traditionally conceived as ‘safe spaces’ for testing new technologies under experimental regulatory supervision, sandboxes offer more than conditional regulatory relief. In the context of AI research, their real potential lies in enabling legal, ethical, policy and technological actors to engage in and pioneer real-time governance experiments. Far from being stopgaps for legal uncertainty, sandboxes institutionalize learning, both legal and ethical, by allowing situated, iterative norm formation together with technical development. Such regulatory AI sandbox reconfigures governance from a paradigm of procedural oversight to one of responsible epistemic co-production (Seferi, 2025). Rather than serving as a temporary regulatory exception, the sandbox becomes a continuous, institutionally embedded learning space for situated ethical inquiry and iterative norm development. It is conceived as a distributed and integrative research infrastructure, a ‘playground’ (Resnick, 2017) in the constructive sense, where researchers from different disciplines and domains, ethicists, legal scholars, and societal stakeholders can collaboratively engage with the uncertainties and contested values of AI systems (Undheim et al., 2022).

5.2.2 The Role of the Research Ethics Committee

The governance of the co-constructive research and innovation environment calls for a redefinition of the role of REC. Rather than bypassing REC, sandboxing invites a reconfiguration of the REC´s tasks: from reviewing individual AI experiments (risk category 3, section 5.1) to overseeing the sandbox as a meta-level governance framework. This model empowers REC to evaluate the quality, reflexivity, and

accountability mechanisms of the sandbox itself, without constraining research within premature or inadequate normative templates. An intra-university responsible AI sandbox could function like a specialized REC sub-committee. This shift supports a holistic, multidisciplinary perspective and concentrates AI expertise within a reflexive, iterative, and ethics-by-design research process environment. Serving as a gateway for external collaboration, the sandbox enables pre- and post-approval engagement while embedding discursive reflection within an established normative framework.

The internal university governance setup, comprising an REC and a responsible AI sandbox, can draw valuable insights from legal/regulatory sandbox models, despite the absence of standardization. Three distinct models emerge: a narrow model focused on product testing (e.g., automated driving); a broad model aimed at regulatory experimentation (e.g., frameworks for AI regulatory sandboxing, Art. 57 AI Act); and hybrid forms enabling context-specific testing of anticipated regulatory regimes, as seen in pre-AI Act initiatives in Spain (Bagni and Seferi, 2025). These configurations (Mobilio and Gianelli, 2025) offer design principles transferable to the internal university context: Most notably, the notion of an *explicit carve-out*, i.e. a defined and bounded space for experimental activity, can be mirrored in a sandbox environment sanctioned by the REC. Within this space, delegated authority enables ethical and regulatory experimentation through iterative and co-constructive processes. A sub-committee structure or designated oversight body could exercise *discretionary powers* akin to those of an oversight authority, facilitating context-sensitive governance of AI research. Establishing *risk thresholds* linked to ethical, technical, and societal dimensions permits differentiated levels of oversight, aligning with REC concerns while preserving room for innovation. *Evaluation and reporting* mechanisms institutionalize reflexivity and ensure accountability over time, supporting a shift from static approvals to dynamic governance.

Crucially, the university sandbox, like its regulatory counterparts, can function as a site of evidence-based regulatory learning. By creating structured input for internal governance and potentially informing external norms, the sandbox helps bridge experimental research and anticipatory regulation (Morgan, 2023). This allows the REC and responsible AI sandbox to evolve beyond compliance gatekeeping towards a model of anticipatory, participatory, and reflexive oversight. Drawing on the AI Act's guiding principles, as well as complementary frameworks such as the EU High-Level Expert Group's (HLEG) ethics guidelines for trustworthy AI (HLEG, 2019), we propose that sandbox environments operationalize 'responsibility' through embedded assessment criteria and metrics (Díaz-Rodríguez et al., 2023). We propose a conceptual structure for capturing these dimensions across the lifecycle of AI systems (section 5.3). To ensure the normative legitimacy of the RAIS, ethical governance must, of course, rest not only on open-ended deliberation but also on procedural clarity, inclusiveness, and transparent communication of expectations and assessment criteria. The AI Act with its risk- and

fundamental-rights-based approach and the HLEG´s guidelines can, at least in the EU, serve as a normative compass in this regard.

5.2.3 Operationalizing Responsible AI by Embedding Normative Reflexivity

The following operationalization of responsible AI for research ethics review purposes is combined with an analysis of regulatory sandboxes as tools for moral imagination (Undheim et al., 2023; Resseguier, 2024) and operates within a framework for responsible innovation, developed in the governance of emergent technologies (Stilgoe et al., 2020). By synthesizing these perspectives, we aim to devise a model for intra-university engagement with AI that is scientifically and pedagogically generative as well as ethically responsive and reflexively designed. Regulatory sandboxes serve a function beyond the minimization of (legal) risk or the facilitation of technological deployment but are valuable precisely because they enable moral imagination under conditions of 'true uncertainty', scenarios in which outcomes cannot be reliably predicted or calculated (Undheim et al., 2023). Within such spaces, the focus shifts from managing known risks to cultivating anticipatory and adaptive forms of governance. This orientation is particularly relevant in the context of AI, where systems increasingly interact with complex social environments and generate outcomes that escape straightforward evaluation. The collaborative, interdisciplinary, and iterative processes within sandboxes emphasize the role of co-learning among regulators, developers, and affected publics. While regulatory sandboxes are typically situated within governmental or market-facing institutions, the core insight that ethical AI development requires structured spaces for open-ended exploration (Francis, 2025) can be usefully transposed into the academic setting. A university-based responsible AI sandbox would not replicate the function of a regulatory sandbox in a narrow sense but would instead adapt its enabling logic to support transdisciplinary learning and responsible design practices. Such sandbox derives its legitimacy from its institutional role in enabling deliberation, capacity-building, and critical inquiry, making it well-suited to address the educational dimensions of responsible AI. By involving students and early-career researchers in real-world projects under conditions of structured uncertainty, the sandbox offers a hands-on context for cultivating ethical sensitivity, interdisciplinary literacy, and design reflexivity. Moreover, the sandbox can serve as a platform for developing soft-law instruments such as codes of conduct, evaluation protocols, or value-sensitive design guidelines that extend beyond individual projects and contribute to the institutional culture of responsible AI development.

The reflexive process can (and should) draw on approaches to algorithmic impact assessment (Selbst, 2021), such as the Fundamental Rights and Algorithms Impact Assessment (Gerards et al., 2022), which provide structured methodologies for identifying and mitigating human rights impacts throughout the algorithmic development process. Integrating such approaches into the framework and assessment practice

strengthens its ability to operationalize fundamental rights considerations, linking reflexive ethical deliberation with concrete, legally informed assessment practices. This alignment would also enhance the approach´s accountability dimension by situating ethical reflection within broader societal and legal frameworks of rights protection and participatory evaluation.

In addition to traditional ethical concerns, responsible AI governance must also address questions of sustainability, encompassing the environmental impact of computationally intensive methods, the infrastructural dependencies of AI research, and issues of digital sovereignty. Particularly in academic settings, where AI experimentation often relies on energy-intensive processes and third-party cloud infrastructures, ethical reflection should extend to the ecological footprint and long-term viability of research practices.

5.3 Approach to a Responsible AI Sandbox - a Focus on Metrics

The technical perspective of the operationalization of responsible AI within a sandbox environment requires a technically grounded and multidimensional algorithmic impact assessment framework. Its purpose is to enhance deliberation with technological means (Mauri et al., 2024). We propose a categorization of assessment dimensions that, at the same time, prepare the ground for ethics-by-design methodologies

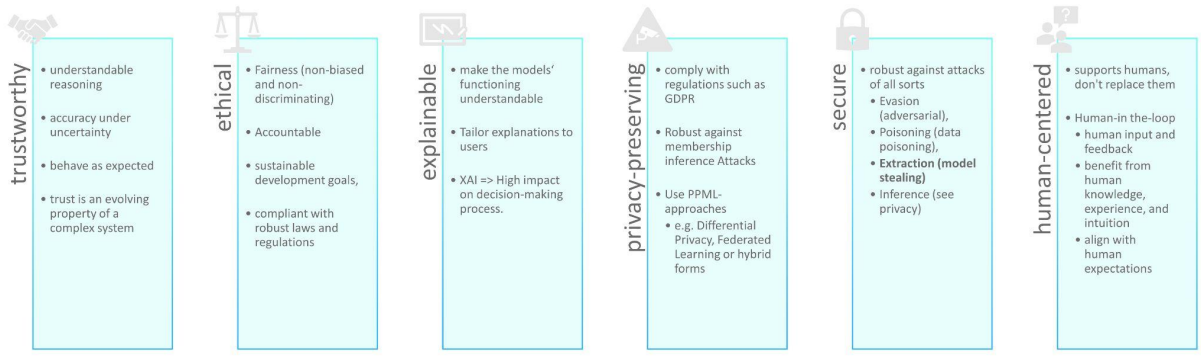

Figure 1: Responsible AI metrics categorization, based on Göllner et al., 2024b.

(Brey and Dainow, 2023) by enabling evaluation metrics in pillars as shown in Fig. 1 (based on Göllner et al., 2024b; High-Level Expert Group on AI, 2019). These pillars function as a guideline or normative-technical blueprint for design, evaluation, and governance of AI systems and embed responsible AI into research and innovation development:

Trustworthiness: AI systems should demonstrate understandable reasoning under uncertainty, maintain accuracy and functional reliability across varying input conditions, and behave consistently with expectations. Possible metrics under this category include calibration error, uncertainty quantification, and out-of-distribution detection rates. In sandbox settings, these metrics are evaluated continuously across model updates to

monitor trust degradation or improvement over time. Trust is not static but evolves as a system interacts with complex environments. It must be treated not as a binary property, but as a dynamic signal within the model lifecycle.

Ethical Alignment: The ethical (in the narrower sense) assessment focuses on fairness, non-discrimination, and accountability. Technical fairness metrics, such as demographic parity, equalized odds, and disparate impact, are computed over protected attributes. Bias mitigation strategies, such as reweighting, adversarial debiasing, or fair representation learning, can be integrated into the sandbox's evaluation logic. Accountability further demands traceability, supported, e.g., through lineage logging. This aligns with the Ethics-by-design approach, focusing on integration of ethical considerations into AI system development from the start (Brey and Dainow, 2024).

Explainability: The decision-making processes of AI systems should be interpretable. Explainable AI (XAI) methods can convey an idea about the decision-making process and support transparency of AI systems. For the proper interpretation of the explanations, domain/expert knowledge is required. Explainability is addressed through both post-hoc and model-intrinsic approaches. A sandbox should support post-hoc explainability via model-agnostic methods (e.g., LIME, SHAP) and model-specific techniques (e.g., Integrated Gradients). Quantitative evaluation dimensions include faithfulness (e.g., input perturbation tests), monotonicity (whether feature importances correlate with performance), sparsity, and explanation stability under perturbations. Explainability-by-design approaches, such as attention-based architectures or self-explaining models, should be benchmarked for interpretability scores. Domain knowledge is necessary for contextual evaluation of explanation plausibility.

Privacy Preservation: Robustness against privacy leakage is essential, especially when handling sensitive or personal data. Technical evaluation includes membership inference attacks, where the area under the ROC curve (Receiver Operating Characteristic) serves as an indicator for memorization risk. Privacy-preserving machine learning techniques, such as differential privacy, federated learning, or secure multiparty computation, must be incorporated where applicable. Privacy guarantees should be auditable, with the sandbox providing reproducible attack simulations and quantitative leakage indicators.

Safety and Security: AI systems should be resilient to a wide spectrum of attacks, including evasion (adversarial examples), data poisoning, model extraction, and inference. AI systems deployed in dynamic environments should be therefore robust to input perturbations. Security assessments include adversarial robustness, model extraction resilience, and poisoning attack tolerance. Robustness is measured under different threat models using standardized frameworks, employing metrics such as worst-case accuracy under bounded perturbations. Model extraction and inversion risk are assessed via black box querying strategies. Poisoning robustness involves training set sanitization efficacy and resilience of performance under perturbed training distributions.

Human-centeredness: Human-in/on/beyond-the-loop architectures ensure that human judgment, experience, and feedback remain integral, aligning AI behaviour with human expectations and societal values. This assessment pillar emphasizes the role of the human not only as an end-user but as an epistemic agent. Human-in-the-loop setups should be formally integrated into the model evaluation cycle, enabling structured user feedback loops, active learning setups, or override mechanisms. Metrics in this category include task performance with vs. without human correction, agreement scores between model and human judgment, and usability or cognitive load assessments. Models must align with human expectations and support decision augmentation, not replacement.

From the technical perspective, metrics-based evaluation is a necessary component for operationalizing responsible AI. Metrics are quantitative indicators which provide measurable criteria for the system behaviour. Yet quantitative metrics alone cannot fully capture the reliability and compliance across all dimensions of AI systems. Responsible AI involves context sensitivity, trade-offs (e.g., between transparency and security) and such aspects as trust or human-centeredness. Many decisions regarding AI depend on domain knowledge, interpretation or ethical judgment. Therefore, it is essential to engage experts and users not only to interpret metric results, but especially to identify limitations and make development decisions. The expert input should be based on interdisciplinary knowledge and integrated directly into the evaluation process. Metrics help structure the evaluation of AI output, behaviour and risks. They also support systematic review that can be used by ethics committees to assess AI systems (Jordan, 2019). The VERIFAI Framework (Göllner and Tropmann-Frick, 2023) implements a large part of the metrics for AI classification models and thereby provides a foundation for the further technical development of comprehensive responsibility assessments. Its modular structure and initial metric coverage enable systematic integration of fairness, explainability, robustness, and privacy evaluations into the model development lifecycle. Given the breadth and heterogeneity of responsibility dimensions, a single monolithic tool is insufficient. Instead, a framework suite approach is better suited, allowing the flexible combination of specialized modules to address domain-specific requirements and to adapt metric application across different AI system types and application contexts.

6 Limitations

The implementation of reflexive ethics governance models such as RAIS must be understood against the backdrop of institutional constraints, as many REC in academic settings remain poorly funded and often lack the interdisciplinary expertise required for AI-related review. To ensure the practical viability of such frameworks, sustainable resourcing will be essential, possibly including dedicated funding lines, specialized staff positions, and closer integration with existing research infrastructure and support units.

Without such structural reinforcement, there is a risk that RAIS-like mechanisms could inadvertently add procedural complexity rather than strengthening ethical reflexivity.

The effectiveness of a reflexive environment such as the RAIS depends not only on its design but also on its alignment with prevailing academic incentive structures, which often prioritize competition, productivity, and intellectual ownership over deliberation and collective reflection. These pressures, combined with hierarchical dynamics within research teams, can discourage open discussion of ethical challenges, particularly among junior researchers or those in precarious positions. To counteract such effects, RAIS-like initiatives should be embedded within institutional frameworks that promote open science, ensure protection for critical participation, and establish participatory governance mechanisms that empower all members of the research community to engage safely and meaningfully in ethical dialogue.

7 Conclusion and Outlook

This paper has identified key structural limitations and epistemic mismatches that constrain REC in their capacity to adequately oversee AI research, particularly where uncertainty, iteration, and socio-technical entanglement are central features. By analogical reasoning from regulatory sandbox models in EU and national jurisdictions and by drawing on a comprehensive responsible AI framework, we have conceptualized the *responsible AI sandbox* (RAIS) as an institutional innovation capable of embedding ethical and legal norms *within* the research process rather than applying them ex ante or ex post.

RAIS functions not merely as a procedural alternative but as a co-constructive governance environment, where responsibility, fairness, risk-mitigation, and trustworthiness are shaped through bounded experimentation, iterative feedback, and situated reflexivity. Unlike traditional front-loaded ethics review, the sandbox allows for dynamic norm development aligned with the unfolding nature of AI technologies. In this model, the REC shifts from a gatekeeping role to one of conditional delegation and continuous oversight, enabling ethics to travel with the research. Such a transformation reframes university governance as an anticipatory, learning-oriented infrastructure bridging innovation and accountability.

The responsible AI sandbox is a proposed institutional infrastructure that embeds ethical reflexivity into AI research and innovation processes within universities. It addresses a ‘grey zone’ of AI research that does not clearly fall under existing REC ethics approval procedures but nonetheless raises emergent normative concerns. A differentiated review approach distinguishes foundational AI research (which may be exempt from REC review) from applied or high-risk projects (which require standard oversight), with the

sandbox offering a governance solution for projects in between, as well as a place for continuous research project reflection after formal approval.

RAIS introduces a metrics-based framework for assessing responsible AI, structured around six pillars: trustworthiness, ethical alignment, explainability, privacy preservation, safety/security, and human-centeredness. These dimensions enable ongoing evaluation across the AI lifecycle using technical and normative indicators, such as calibration error, fairness metrics, privacy leakage risks, adversarial robustness, and interpretability scores. Further, the RAIS framework can serve as a site for integrating sustainability considerations into ethical deliberation, promoting awareness of resource use, infrastructural resilience, and institutional autonomy as essential dimensions of responsible research on and with AI.

Inspired by regulatory sandboxes in law and technology, RAIS operationalizes ‘ethical co-construction’ through iterative, multidisciplinary collaboration. Unlike traditional oversight focused on compliance, the sandbox supports reflexive, adaptive governance, where researchers, ethicists, legal scholars, and societal actors collaboratively shape norms in real-time. This positions the REC not as a gatekeeper but as an institutional steward monitoring the sandbox’s integrity, transparency, and learning capacity.

Rather than approving individual projects, an embedded REC sub-committee can grant environment-level approval, enabling supervised ethical experimentation within a bounded domain. RAIS thus functions as a site of anticipatory regulation and institutional learning, generating soft-law instruments (e.g., codes of conduct) and ethical infrastructures. It draws legitimacy from its capacity to enable and facilitate deliberation, build capacity, and cultivate interdisciplinary responsibility.

Acknowledgements

This research was made possible through funding by the German Federal Ministry of Research, Technology and Space (Bundesministerium für Forschung, Technologie und Raumfahrt) as part of the Responsible Advanced Intelligent Methodologies and Skills Lab (R-AIMS) project at Hamburg University of Applied Sciences. The authors gratefully acknowledge this support, which made the present contribution possible. The authors would also like to thank the reviewers for the valuable comments. The views expressed are solely those of the authors.

Conflicts of interest

The authors declare no conflicts of interest.

References

d'Aquin, M., Troullinou, P., O'Connor, N. E., Cullen, A., Faller, G., & Holden, L. (2018, December). Towards an' ethics by design' methodology for AI research projects. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (pp. 54-59).

Bagni F. and Seferi F. (eds.) (2025), Regulatory sandboxes for AI and Cybersecurity. Questions and answers for stakeholders. CINI's Cybersecurity National Lab. ISBN: 9788894137378.

Bouhouita-Guermech, S., Gogognon, P., & Bélisle-Pipon, J. C. (2023). Specific challenges posed by artificial intelligence in research ethics. Frontiers in artificial intelligence, 6, 1149082.

Brenneis, A., Gehring, P., & Lamadé, A. (2024). Zwischen fachlichen Standards und wilder Innovation: Zur Begutachtung von Big Data-und KI-Projekten in Forschungs-Ethikkommissionen. Ethik in der Medizin, 1-19.

Brenneis, A., & Burden, T. (2025). Meeting report: 'Challenges Posed by AI for the Work of Research Ethics Committees'. Conference, 2024, Hannover, DE. TATuP-Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis, 34(1), 70-71.

Brey, P., & Dainow, B. (2024). Ethics by design for artificial intelligence. AI and Ethics, 4(4), 1265-1277.

Centre Responsible Digitality (ZEVEDI): Research Ethics for AI Research Projects. Guidelines to Support the Work of Ethics Committees at Universities, Darmstadt 2023, 19 pp.

Calo, R. (2015). Robotics and the Lessons of Cyberlaw. Calif. L. Rev., 103, 513.

Crootof, R., & Ard, B. J. (2020). Structuring techlaw. Harv. JL & Tech., 34, 347.

Feindt, P. H., & Weiland, S. (2018). Reflexive governance: exploring the concept and assessing its critical potential for sustainable development. Introduction to the special issue. Journal of Environmental Policy & Planning, 20(6), 661–674. https://doi.org/10.1080/1523908X.2018.1532562

Ferretti, A., Ienca, M., Sheehan, M., Blasimme, A., Dove, E. S., Farsides, B. & Vayena, E. (2021). Ethics review of big data research: What should stay and what should be reformed?. BMC medical ethics, 22(1), 51.

Francis, K. (2025). The need for an ethical approach to regulatory sandboxes. In: Bagni F. and Seferi F. (eds.) (2025), Regulatory sandboxes for AI and Cybersecurity. Questions and answers for stakeholders. CINI's Cybersecurity National Lab.

Gerards, J., Schäfer, M. T., Muis, I., & Vankan, A. (2022). Fundamental rights and algorithms impact assessment (fraia).

Gille, M. & Tropmann-Frick, M. & Schomacker, T. (2024). Balancing public interest, fundamental rights, and innovation: The EU's governance model for non-high-risk AI systems. Internet Policy Review, 13(3).

Göllner, S., Tropmann-Frick, M., & Brumen, B. (2024a). Towards a Definition of a Responsible Artificial Intelligence. In Information Modelling and Knowledge Bases XXXV (pp. 40-56). IOS Press.

Goellner, S., Tropmann-Frick, M., & Brumen, B. (2024b). Responsible Artificial Intelligence: A Structured Literature Review. arXiv preprint arXiv:2403.06910.

Göllner, S., & Tropmann-Frick, M. (2023). VERIFAI-A Step Towards Evaluating the Responsibility of AI-Systems. In BTW 2023 (pp. 933-941).

González-Esteban, E. & Patrici, C. (2022). Ethically governing artificial intelligence in the field of scien.fic research and innovation. Heliyon, 8.

Hadley, E., Blatecky, A. & Comfort, M. Investigating algorithm review boards for organizational responsible artificial intelligence governance. AI Ethics 5, 2485–2495 (2025). https://doi.org/10.1007/s43681-024-00574-8

Hagendorff T (2020) The ethics of AI ethics. An evaluation of guidelines. Minds and Machines 30: 99–120. Crossref.

High-Level Expert Group on AI (HLEG) (2019) Ethics guidelines for trustworthy AI. Available at: https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html.

Jordan, S.R. (2019). Designing Artificial Intelligence Review Boards: Creating Risk Metrics for Review of AI, 2019 IEEE International Symposium on Technology and Society (ISTAS), 2019, pp. 1-7, doi: 10.1109/ISTAS48451.2019.8937942.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature machine intelligence, 1(9), 389-399.

Jones, M. L. (2018). Does technology drive law? The dilemma of technological exceptionalism in cyberlaw. U. Ill. JL Tech. & Pol'y, 249.

König, H., Baumann, M. F., & Coenen, C. (2021). Emerging technologies and innovation—hopes for and obstacles to inclusive societal co-construction. Sustainability, 13(23), 13197.

Lenzini, G. (2025). Artificial Intelligence Ethics: Challenges for a Computer Science Ethics Board with a Focus on Autonomy. In The Routledge Handbook of Artificial Intelligence and International Relations (pp. 382-391). Routledge.

Mantelero, A. (2024). The Fundamental Rights Impact Assessment (FRIA) in the AI Act: Roots, legal obligations and key elements for a model template. Computer Law & Security Review, 54, 106020.

Masso, A., Gerassimenko, J., Kasapoglu, T., & Beilmann, M. (2025). Research Ethics Committees as Knowledge Gatekeepers: The Impact of Emerging Technologies on Social Science Research. Journal of Responsible Technology, 100112.

Mauri, A., Hsu, Y. C., Verma, H., Tocchetti, A., Brambilla, M., & Bozzon, A. (2024). Policy Sandboxing: Empathy As An Enabler Towards Inclusive Policy-Making. Proceedings of the ACM on Human-Computer Interaction, 8(CSCW2), 1-42.

Mittelstadt, B. D. (2022). Principles alone cannot guarantee ethical AI. Nature Machine Intelligence, 4(1), 5–7.

Mobilio, G. and Gianelli, M. (2025). In: Bagni F. and Seferi F. (eds.) (2025), Regulatory sandboxes for AI and Cybersecurity. Questions and answers for stakeholders. CINI's Cybersecurity National Lab.

Morgan, D. (2023, August). Anticipatory regulatory instruments for ai systems: A comparative study of regulatory sandbox schemes. In Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (pp. 980-981).

Petermann, M., Tempini, N., Kherroubi Garcia, I., Whitaker, K., & Strait, A. (2022). Looking before we leap: Expanding ethical review processes for AI and data science research.

Plato-Shinar, R., & Godwin, A. (2025). Regulatory Cooperation in AI Sandboxes: Insights from Fintech.

Ranchordás, S. (2021). Experimental Regulations for AI: Sandboxes for Morals and Mores. Morals & Machines, 1(1), 86-100.

Rangoni, B. (2022). Experimentalist governance. In Handbook on Theories of Governance (pp. 592-603). Edward Elgar Publishing.

Resseguier, A. and Ufert, F. (2023). AI research ethics is in its infancy: the EU's AI Act can make it a grown-up. Research Ethics, 20(2), 143-155. https://doi.org/10.1177/17470161231220946 (Original work published 2024)

Resseguier, A. (2024). Research ethics frameworks for artificial intelligence: The twofold need for compliance requirements and for an open process of reflection and attention. In Smart Ethics in the Digital World: Proceedings of the ETHICOMP 2024. 21th International Conference on the Ethical and Social Impacts of ICT (pp. 122-124). Universidad de La Rioja.

Reijers, W., Wright, D., Brey, P., Weber, K., Rodrigues, R., O'Sullivan, D., & Gordijn, B. (2018). Methods for practising ethics in research and innovation: A literature review, critical analysis and recommendations. Science and engineering ethics, 24, 1437-1481.

Resnick, M. (2017). Lifelong kindergarten: Cultivating creativity through projects, passion, peers, and play. MIT press.

Sabel, C. F., & Zeitlin, J. (2012). Experimentalist governance. In Levi-Faur, D. (Ed.), Oxford Handbook of Governance. Oxford University Press.

Seferi, F. (2025). A comparative analysis of regulatory sandboxes from selected use cases: Insights from recurring operational practices. In Regulatory sandboxes for AI and Cybersecurity. Questions and answers for stakeholders (pp. 145-176). CINI's Cybersecurity National Lab.

Selbst, A. D. (2021). An institutional view of algorithmic impact assessments. Harv. JL & Tech., 35, 117.

Stone, D. (2012). Transfer and translation of policy. Policy studies, 33(6), 483-499.

Voß, J. P., & Kemp, R. (2006). Sustainability and reflexive government: introduction. In Reflexive governance for sustainable development. Edward Elgar Publishing.

Zimmer, M. (2018). Addressing conceptual gaps in big data research ethics: An application of contextual integrity. Social Media+ Society, 4(2), 2056305118768300.