**DOI https://doi.org/10.30525/2592-8813-2026-2-38**

# BEYOND COMPLIANCE: EMBEDDING ETHICS-BY-DESIGN IN ARTIFICIAL INTELLIGENCE SYSTEMS UNDER THE EU AI ACT

***Tamerlan Hajiyev,***
*PhD, Lecturer at the Department of International Relations and Foreign Politics,*
*Academy of Public Administration under the President of the Republic of Azerbaijan*
*(Baku, Azerbaijan)*
*ORCID ID: 0009-0008-1268-1954*

**Abstract.** This article explores the evolution of artificial intelligence governance in the European Union, focusing on the shift from regulatory compliance toward ethics-by-design. While the EU Artificial Intelligence Act establishes a comprehensive risk-based framework, the study argues that compliance alone is insufficient to address deeper ethical challenges related to human autonomy, dignity, fairness, and power asymmetries in algorithmic systems. The article conceptualizes ethics-by-design as a form of operational normativity that embeds ethical principles directly into the lifecycle of AI systems. It identifies key mechanisms for such integration, including ethical impact assessments, algorithmic auditing, transparency tools, and human oversight. At the same time, it highlights challenges such as translating abstract values into technical practices and avoiding superficial formalization. The study concludes that ethics-by-design is a necessary condition for the legitimacy of AI governance in the EU. By integrating ethical reasoning into system design, it becomes possible to bridge the gap between legal compliance and substantive justice, reinforcing a human-centric model of digital governance.

**Key words:** Artificial Intelligence Governance, Ethics-by-Design, EU AI Act, Trustworthy AI, Human-Centric AI, Algorithmic Accountability, Risk-Based Regulation, Digital Ethics, AI Regulation, European Union.

**Introduction.** The regulation of artificial intelligence has rapidly evolved from a predominantly technical and sector-specific concern into one of the central issues of contemporary legal, political, and ethical theory. This transformation is driven not only by the accelerated deployment of algorithmic systems across domains such as public administration, healthcare, finance, education, and security, but also by the increasingly profound influence these systems exert on the fundamental conditions of human existence. Artificial intelligence no longer functions merely as an auxiliary tool; rather, it actively shapes decision-making environments, redistributes power, and redefines the boundaries of human autonomy, dignity, privacy, and equality. In response to these developments, the European Union has positioned itself at the forefront of global AI governance by adopting a comprehensive regulatory framework, most notably the Artificial Intelligence Act, which seeks to establish a coherent model for the development and use of trustworthy and human-centric AI.

At the core of the European regulatory approach lies a tension that defines the contemporary debate on AI governance: the distinction between formal regulatory compliance and substantive ethical legitimacy. Modern legal systems, particularly in technologically dynamic contexts, tend to prioritize operational clarity by introducing categories, thresholds, and procedural requirements designed to ensure predictability and enforceability. The EU AI Act exemplifies this logic through its risk-based classification of AI systems and the corresponding obligations imposed on providers and deployers. While such an approach is indispensable for the functioning of a complex regulatory environment, it simultaneously reveals its own limitations. Compliance with legal requirements does not necessarily guarantee that an AI system is ethically acceptable in a broader normative sense. A system may be lawful, technically robust, and procedurally compliant, yet still undermine human autonomy, reinforce structural inequalities, obscure accountability, or erode meaningful human control over decision-making processes.

This gap between legality and ethical legitimacy has given rise to an increasing recognition that effective AI governance cannot be reduced to ex post regulatory oversight or formal adherence to predefined rules. Instead, there is a growing need to integrate ethical considerations directly into the design, development, and deployment of AI systems. It is within this context that the concept of ethics-by-design emerges as a critical paradigm. Rather than treating ethics as an external evaluative layer applied after a system has been developed, ethics-by-design calls for the proactive incorporation of normative principles into the very architecture of technological systems. This shift reflects a broader transformation in regulatory thinking: from reactive governance focused on mitigating risks, to anticipatory governance aimed at shaping technological trajectories in accordance with fundamental values (Floridi, 2018, p. 697).

The intellectual foundations of this paradigm are deeply embedded in the European tradition of human-centric AI. Early policy frameworks, including the Ethics Guidelines for Trustworthy AI developed by the European Commission's High-Level Expert Group, articulated a vision of AI that is not only lawful but also ethical and socially robust throughout its lifecycle. These guidelines identified key principles such as respect for human agency, transparency, fairness, accountability, and the protection of fundamental rights. However, despite their conceptual significance, such frameworks have often remained at the level of soft law, lacking direct enforceability and practical integration into technological processes. As a result, a persistent disjunction has emerged between the articulation of ethical principles and their actual implementation within AI systems.

The adoption of the EU AI Act represents a significant step toward bridging this gap by introducing binding legal obligations and a structured risk-based regulatory model. The Act distinguishes between prohibited practices, high-risk systems, limited-risk applications, and minimal-risk uses, thereby tailoring regulatory intensity to the potential impact of the technology. Nevertheless, the effectiveness of this model ultimately depends on how compliance is conceptualized. If compliance is treated as the ultimate objective, AI governance risks devolving into a procedural exercise that prioritizes formal adherence over substantive outcomes. In such a scenario, ethical considerations remain secondary, contingent upon voluntary initiatives rather than embedded requirements. Conversely, if compliance is understood as a minimum threshold rather than a sufficient condition, it becomes possible to reconceptualize AI regulation as a framework for institutionalizing ethical values within technological systems (Jobin, 2019, p. 375).

This reconceptualization is particularly important given the transformative nature of artificial intelligence. AI systems do not merely execute predefined tasks; they participate in structuring social reality by influencing how information is processed, how decisions are made, and how individuals are categorized and evaluated. In doing so, they introduce new forms of opacity, complexity, and distributed responsibility that challenge traditional legal doctrines. Harm or injustice may arise not from a single identifiable action, but from the cumulative effects of design choices, data selection, model optimization, and contextual deployment. Consequently, ethical considerations cannot be confined to retrospective assessments of outcomes; they must be integrated into every stage of the AI lifecycle, from initial design to continuous monitoring.

Against this backdrop, the central research question of this article can be formulated as follows: to what extent can the European Union's current regulatory framework, particularly the AI Act, move beyond a compliance-based approach and facilitate the genuine integration of ethics-by-design as a structural component of AI governance? This question is not merely technical; it reflects a broader normative choice between two competing visions of the digital future. One vision treats ethics as an optional supplement to regulation, valuable for enhancing trust and reputation but ultimately external to the core functioning of AI systems. The other envisions ethics as an intrinsic element of technological design, without which legal regulation remains insufficient and perpetually reactive.

The aim of this article is to argue in favor of the latter perspective. It seeks to demonstrate that sustainable and legitimate AI governance in the European context requires a shift from formal compliance to substantive ethical integration. In this framework, law is not abandoned but reinterpreted as a mechanism for institutionalizing ethical consensus rather than merely enforcing minimal standards. The article pursues three interrelated objectives: first, to reconstruct the theoretical foundations of ethics-by-design within the European normative tradition; second, to critically assess the limitations of the risk-based regulatory approach embodied in the AI Act; and third, to propose a conceptual model for embedding ethical principles into the lifecycle of AI systems through concrete governance mechanisms.

Methodologically, the study combines normative legal analysis with philosophical-ethical inquiry and institutional examination. The legal dimension focuses on the interpretation of the EU AI Act and related policy instruments. The ethical analysis explores the conceptual content of key principles such as human dignity, autonomy, fairness, and accountability, emphasizing their role as foundational rather than auxiliary elements of regulation. The institutional perspective, in turn, addresses the mechanisms through which these principles can be operationalized, including impact assessments, algorithmic audits, transparency tools, and human oversight structures. This interdisciplinary approach reflects the complexity of AI governance, which cannot be adequately addressed within the confines of a single discipline.

The novelty of the article lies in its reconceptualization of ethics-by-design as a structural condition for the legitimacy of AI governance rather than a voluntary or supplementary practice. By framing ethical integration as a necessary complement to legal regulation, the study contributes to ongoing debates on the future of European digital governance and the possibility of aligning technological innovation with the preservation of fundamental human values. Ultimately, the transition beyond compliance is presented not as a rejection of law, but as its deepening through the incorporation of ethical reasoning into the very fabric of technological development.

**Discussion.** The findings of this study demonstrate that the European approach to artificial intelligence governance is moving beyond traditional regulatory compliance toward a more integrated ethical model. Although the EU AI Act provides one of the most comprehensive legal frameworks for AI regulation, the analysis suggests that compliance alone is insufficient to address broader ethical concerns related to human autonomy, fairness, accountability, and the societal impact of algorithmic systems.

The study confirms arguments advanced in existing scholarship that principle-based AI ethics remains limited when ethical values are not effectively translated into technological and institutional practices. In this context, ethics-by-design emerges as a significant governance paradigm because it seeks to embed ethical reasoning directly into the lifecycle of AI systems rather than treating ethics as an external evaluative mechanism. The article demonstrates that tools such as ethical impact assessments, algorithmic auditing, transparency measures, and human oversight structures can play an important role in operationalizing ethical principles within AI development and deployment.

At the same time, the research highlights several important challenges. One of the key difficulties lies in translating abstract ethical concepts, such as fairness or autonomy, into concrete technical standards and measurable procedures. Excessive formalization may reduce ethics to procedural checklists, weakening its critical and reflective function. In addition, the effectiveness of ethics-by-design depends heavily on institutional culture, interdisciplinary cooperation, and the availability of organizational resources.

The analysis also underlines the broader normative significance of the European model of AI governance. The EU's emphasis on human-centric and trustworthy AI reflects an attempt to align technological innovation with democratic values and fundamental rights. In this regard, ethics-by-design strengthens the legitimacy of AI governance by bridging the gap between legal compliance and substantive ethical responsibility.

The practical significance of the study lies in its contribution to ongoing discussions on the future of AI governance. The proposed framework may support policymakers, regulatory institutions, and developers in designing AI systems that are not only legally compliant, but also socially responsible and ethically sustainable. Ultimately, the article argues that the long-term legitimacy of AI governance in the European Union will depend on the successful integration of ethical principles into the core architecture of artificial intelligence systems.

**Theoretical Foundations of Ethics-by-Design: From Principles to Operational Norms**

The concept of ethics-by-design has emerged as a response to a fundamental limitation in contemporary approaches to artificial intelligence governance: the persistent gap between the articulation of ethical principles and their effective implementation within technological systems. While the proliferation of ethical guidelines, frameworks, and declarations has significantly enriched the normative discourse surrounding artificial intelligence, these instruments have often remained insufficiently connected to the practical realities of system design and deployment. As a result, ethics has frequently been positioned as an external evaluative layer – invoked in policy rhetoric but only loosely integrated into the operational logic of AI systems. The ethics-by-design paradigm seeks to overcome this limitation by reconfiguring ethics as an intrinsic component of technological architecture rather than a post hoc regulatory consideration.

At its theoretical core, ethics-by-design is grounded in a rethinking of the relationship between normativity and technology. Traditional regulatory paradigms have tended to treat technology as a neutral object subject to external control through legal rules and institutional oversight. In this view, ethical evaluation occurs after the fact, assessing whether a given system complies with predefined standards or produces acceptable outcomes. However, such an approach becomes increasingly inadequate in the context of artificial intelligence, where systems are not merely tools but dynamic, adaptive, and often opaque infrastructures that shape the conditions under which decisions are made. The shift toward ethics-by-design reflects the recognition that normativity must be embedded within the design process itself, influencing not only what systems do, but how they are conceptualized, structured, and integrated into social environments (Floridi, 2019).

This shift can be understood as part of a broader transformation in normative theory, moving from principle-based ethics toward what may be described as operational normativity. Classical ethical frameworks – whether grounded in deontological commitments to duty and rights, consequentialist evaluations of outcomes, or virtue-based considerations of character – have historically provided abstract criteria for moral judgment. While these traditions remain indispensable as sources of normative orientation, their direct application to complex technological systems presents significant challenges. Artificial intelligence operates through layers of data processing, model training, optimization functions, and interface design, each of which introduces specific points of ethical relevance that cannot be adequately addressed through high-level principles alone. Ethics-by-design therefore requires the translation of abstract values into context-sensitive, technically actionable norms that can guide decision-making at each stage of the system lifecycle.

Within the European intellectual and regulatory tradition, this process of translation has been particularly visible in the development of the concept of trustworthy AI**.** The emphasis on trustworthiness reflects an attempt to bridge the gap between ethical aspiration and institutional practice by articulating a set of requirements that are both normatively grounded and practically applicable. Key principles such as respect for human dignity, autonomy, fairness, accountability, and transparency are not treated as isolated values, but as interconnected elements of a broader normative architecture. However, the significance of these principles lies not in their declarative formulation, but in their capacity to inform concrete design choices. For instance, the principle of autonomy implies not only the avoidance of coercive or manipulative systems, but also the active design of interfaces that support meaningful human agency and informed decision-making. Similarly, the principle of fairness

extends beyond the prohibition of discrimination to encompass the careful curation of training data, the selection of appropriate evaluation metrics, and the continuous monitoring of system outputs for unintended biases.

The transition from principles to operational norms also requires a redefinition of responsibility in the context of AI systems. In traditional legal frameworks, responsibility is often associated with identifiable actors and discrete actions. By contrast, artificial intelligence systems distribute agency across a network of designers, developers, data providers, deployers, and users, as well as across the technical components of the system itself. This distributed nature of agency complicates the attribution of responsibility and raises questions about how ethical obligations can be effectively enforced. Ethics-by-design addresses this challenge by embedding responsibility into the structure of the system, ensuring that accountability is not merely a matter of ex post attribution but is reflected in the design of processes, documentation practices, and oversight mechanisms. In this sense, responsibility becomes a property of the system's architecture rather than solely a characteristic of individual actors (Watcher, 2018, p. 862).

Another crucial dimension of ethics-by-design is its temporal orientation. Conventional regulatory approaches often operate retrospectively, responding to harms after they have occurred. In contrast, ethics-by-design is inherently anticipatory**,** seeking to identify and mitigate ethical risks before they materialize. This anticipatory logic aligns with broader developments in governance theory, particularly in the context of emerging technologies, where uncertainty and complexity necessitate proactive forms of regulation. By integrating ethical considerations into early stages of system development – such as problem definition, data selection, and model design – ethics-by-design aims to prevent the entrenchment of harmful patterns that may be difficult or impossible to correct at later stages.

At the same time, the operationalization of ethical principles introduces its own set of challenges. One of the central difficulties lies in the potential tension between competing values. For example, efforts to enhance transparency may conflict with the protection of intellectual property or privacy, while the pursuit of fairness may require trade-offs between different definitions of equality. Ethics-by-design does not eliminate these tensions, but it provides a structured framework for addressing them through explicit design choices and documented reasoning processes. In this regard, ethical decision-making becomes an integral part of system engineering, requiring interdisciplinary collaboration between legal experts, ethicists, engineers, and domain specialists.

Furthermore, the move toward operational normativity raises important questions about standardization and measurement. In order for ethical principles to be effectively embedded in AI systems, they must be translated into criteria that can be assessed, monitored, and, where necessary, enforced. This has led to the development of tools such as ethical impact assessments, algorithmic auditing frameworks, and standardized documentation practices. While these instruments represent significant progress, they also risk reducing complex ethical considerations to checklists or formal procedures. A key theoretical challenge, therefore, is to ensure that the process of operationalization does not strip ethical principles of their normative depth, but rather preserves their capacity to guide meaningful reflection and critical evaluation.

In this context, ethics-by-design can be understood as a form of normative mediation between abstract values and concrete technological practices. It does not seek to replace traditional ethical theory, but to extend it into new domains by developing mechanisms through which ethical reasoning can inform technical decision-making. This mediation is particularly important in the European context, where the legitimacy of AI governance is closely tied to the protection of fundamental rights and the maintenance of democratic values. By embedding ethics into the design of AI systems, it becomes possible to align technological innovation with the broader normative commitments of the European legal order.

Ultimately, the theoretical significance of ethics-by-design lies in its capacity to reconceptualize the role of ethics in technological governance. Rather than functioning as an external constraint or a secondary consideration, ethics becomes a constitutive element of system design, shaping both the objectives and the methods of AI development. This reconceptualization challenges the traditional separation between technical and normative domains, suggesting that the design of AI systems is inherently a moral and political activity. In doing so, it lays the groundwork for a more integrated approach to AI governance, in which ethical principles are not merely proclaimed, but systematically embedded in the structures that govern technological behavior (Veale, 2021, p. 104).

This theoretical framework provides the foundation for the subsequent analysis of the EU AI Act. While the Act represents a significant advancement in the legal regulation of artificial intelligence, its effectiveness ultimately depends on its capacity to incorporate and operationalize the principles outlined above. The following section therefore turns to a critical examination of the risk-based regulatory model, with particular attention to its strengths, limitations, and potential for enabling – or constraining – the transition toward ethics-by-design.

The EU AI Act and the Limits of Risk-Based Regulation

The adoption of the European Union's Artificial Intelligence Act represents a landmark moment in the global governance of emerging technologies. As the first comprehensive and binding legal framework specifically dedicated to artificial intelligence, the Act reflects the EU's longstanding ambition to position itself as a normative leader in shaping a human-centric and rights-based digital order. At the heart of this regulatory architecture lies the risk-based approach, which classifies AI systems according to the level of potential harm they may pose to individuals and society, and calibrates legal obligations accordingly. This model is widely regarded as both innovative and pragmatic, offering a structured method for balancing technological innovation with the protection of fundamental rights. However, despite its strengths, the risk-based paradigm also reveals significant conceptual and practical limitations when assessed from the perspective of ethics-by-design.

The core logic of the EU AI Act is built upon a hierarchical categorization of AI systems: prohibited practices, high-risk systems, limited-risk applications, and minimal-risk uses. This stratification enables regulators to focus their attention on those systems that are deemed most likely to produce serious harm, particularly in areas such as biometric identification, critical infrastructure, employment, law enforcement, and access to essential services. High-risk systems are subject to extensive requirements, including conformity assessments, risk management procedures, data governance standards, technical documentation, transparency obligations, and human oversight mechanisms. In contrast, systems categorized as presenting minimal or limited risk are subject to lighter regulatory burdens, thereby preserving space for innovation and experimentation.

From a legal and administrative perspective, this approach offers several advantages. It introduces clarity and predictability into the regulatory environment, facilitates enforcement by defining clear categories and obligations, and avoids the pitfalls of overly rigid or one-size-fits-all regulation. Moreover, by explicitly prohibiting certain uses of AI – such as manipulative or exploitative practices that undermine human autonomy – the Act demonstrates a commitment to safeguarding core European values. In this sense, the risk-based model represents a sophisticated attempt to operationalize the principle of proportionality within the domain of AI governance (Stix, 2021, p. 27).

Nevertheless, when examined through the lens of ethical integration, the limitations of this model become increasingly apparent. One of the most fundamental challenges lies in the reduction of ethical complexity to quantifiable risk categories**.** While risk assessment is an essential tool for regulatory decision-making, it inherently prioritizes measurable and foreseeable harms, often at the expense of more diffuse, long-term, or systemic ethical concerns. Issues such as the gradual erosion of human agency, the normalization of algorithmic decision-making, or the subtle reinforcement of social hierarchies may not be easily captured within predefined risk thresholds, yet they have profound implications for the normative structure of society.

Furthermore, the risk-based approach tends to reinforce a compliance-oriented mindset**,** in which the primary objective of developers and deployers is to satisfy regulatory requirements rather than to engage in deeper ethical reflection. In such a framework, ethical considerations risk being instrumentalized as components of risk management rather than recognized as independent normative commitments. The danger here is not that ethical principles are ignored, but that they are translated into procedural checklists that can be formally satisfied without necessarily transforming the underlying logic of system design. As a result, compliance may become a substitute for ethical responsibility, rather than its institutional expression.

Another critical limitation concerns the temporal structure of risk regulation**.** The AI Act, like many regulatory instruments, is primarily oriented toward the identification and mitigation of risks at specific stages of the system lifecycle, particularly prior to market deployment. While this ex ante approach represents a significant improvement over purely reactive models, it remains insufficient in the context of AI systems that evolve over time through continuous learning, data updates, and interaction with dynamic environments. Ethical challenges may emerge not only at the moment of deployment, but also during prolonged use, adaptation, and integration into complex socio-technical systems. A regulatory model that focuses predominantly on initial risk classification may struggle to account for these evolving dimensions.

In addition, the categorization of AI systems into discrete risk levels can obscure the context-dependent nature of ethical evaluation**.** The same technological system may produce vastly different effects depending on how, where, and by whom it is deployed. For example, an algorithm used for resource allocation in a private commercial setting may raise different ethical concerns than a similar system used in public administration or law enforcement. The risk-based model, however, relies on generalized categories that may not fully capture these contextual variations. This creates a tension between the need for regulatory standardization and the inherently situated character of ethical judgment.

The limitations of the risk-based approach also become evident when considering the issue of distributed responsibility**.** The AI Act assigns obligations to specific actors – primarily providers and deployers – thereby maintaining the traditional legal logic of identifiable duty-bearers. However, as noted in the previous section, AI systems are the product of complex networks involving multiple stakeholders, including data suppliers, model developers, system integrators, and end-users. Ethical risks often arise from interactions within this network rather than from the actions of a single actor. By focusing on formal roles within the regulatory framework, the Act may not fully address the diffuse and relational nature of responsibility in AI ecosystems (Morley, 2020, p. 2147).

Moreover, the emphasis on risk mitigation can inadvertently lead to a defensive model of governance, in which the primary goal is to avoid harm rather than to actively promote positive ethical outcomes. While the prevention of harm is undoubtedly essential, it does not exhaust the normative potential of AI governance. A truly human-centric approach would require not only minimizing risks, but also designing systems that enhance human capabilities, support social inclusion, and contribute to the public good. The current structure of the AI Act, however, is more explicitly oriented toward constraint than toward normative innovation, leaving the positive dimension of ethical design underdeveloped.

It is important to note that these limitations do not undermine the significance of the AI Act as a regulatory achievement. On the contrary, the Act provides a necessary legal foundation for the governance of artificial intelligence and establishes a baseline of protection that is indispensable in a rapidly evolving technological landscape. The critique advanced here is not directed at the existence of risk-based regulation as such, but at its insufficiency as a comprehensive model for addressing the ethical dimensions of AI. In other words, the problem is not that the AI Act does too much, but that it cannot, by itself, do enough.

This recognition points toward the need for a complementary framework that extends beyond risk classification and compliance. Ethics-by-design offers such a framework by shifting the focus from external regulation to internal system architecture. Rather than asking only whether a system meets regulatory requirements, ethics-by-design asks how ethical values are embedded within the processes of design, development, and deployment. It emphasizes continuous reflection, interdisciplinary collaboration, and the integration of normative considerations at every stage of the AI lifecycle.

In this light, the relationship between the AI Act and ethics-by-design should not be understood as oppositional, but as structurally interdependent. The Act establishes the legal conditions under which AI systems must operate, while ethics-by-design provides the normative orientation necessary to guide the interpretation and implementation of these conditions. Without the legal framework, ethical principles risk remaining abstract and unenforceable; without ethical integration, legal compliance risks becoming formalistic and insufficient.

The analysis presented in this section thus leads to a central conclusion: the risk-based regulatory model, while indispensable, represents only a partial response to the challenges of AI governance. Its effectiveness ultimately depends on its capacity to be complemented by a deeper integration of ethical principles into the design and operation of AI systems. The next section therefore turns to the practical dimension of this integration, examining the mechanisms, challenges, and governance models through which ethics-by-design can be realized in practice.

**Embedding Ethics into AI Systems: Mechanisms, Challenges, and Governance Models**

If the previous sections have demonstrated the conceptual necessity of moving beyond a purely compliance-based approach and the structural limitations of risk-based regulation, the question that follows is inherently practical: how can ethical principles be effectively embedded into the lifecycle of artificial intelligence systems? The paradigm of ethics-by-design only acquires normative significance if it can be translated into concrete mechanisms capable of influencing technological development in a systematic and verifiable manner. This requires not only the articulation of ethical goals, but also the creation of institutional, technical, and procedural instruments through which these goals can be operationalized.

At the most fundamental level, embedding ethics into AI systems involves a shift from external evaluation to internal integration**.** Rather than treating ethical assessment as a separate stage following system development, ethics-by-design integrates normative considerations into each phase of the AI lifecycle: problem definition, data collection, model development, testing, deployment, and continuous monitoring. This lifecycle-oriented perspective reflects the dynamic nature of AI systems, which evolve over time and interact with complex social environments. Consequently, ethical governance must also be continuous, adaptive, and reflexive (Mittelstadt, 2016, p. 3).

One of the central mechanisms for operationalizing ethics-by-design is the ethical impact assessment (EIA)**.** Building upon existing practices such as data protection impact assessments under the GDPR, EIAs aim to identify potential ethical risks at an early stage of system development. Unlike traditional risk assessments, which focus primarily on measurable harms, ethical impact assessments are broader in scope, encompassing questions of fairness, autonomy, dignity, and societal impact. They require developers and organizations to engage in anticipatory reflection, considering not only how a system functions, but also how it may reshape social relations and institutional practices. Importantly, the effectiveness of EIAs depends on their integration into decision-making processes, rather than their treatment as formal documentation exercises.

Closely related to impact assessments is the practice of algorithmic auditing**,** which provides a mechanism for evaluating the behavior of AI systems both before and after deployment. Auditing can take multiple forms, including internal audits conducted by developers, external audits performed by independent bodies, and continuous monitoring systems that track performance over time. From an ethical perspective, algorithmic auditing serves as a tool for ensuring accountability and transparency,

enabling stakeholders to detect biases, assess compliance with ethical standards, and identify unintended consequences. However, auditing also raises questions about standardization, access to proprietary systems, and the independence of evaluators, highlighting the need for robust governance frameworks that balance transparency with legitimate concerns such as intellectual property and security.

Another key dimension of ethics-by-design is the incorporation of explainability and transparency mechanisms**.** In the context of AI, transparency is not merely a matter of disclosing information, but of ensuring that system behavior can be meaningfully understood by relevant stakeholders. This includes not only technical explanations of model functioning, but also contextual information about data sources, decision criteria, and potential limitations. Explainability is particularly important in high-stakes domains, where individuals are affected by automated decisions and require the ability to challenge or contest outcomes. At the same time, the pursuit of explainability must be carefully balanced against the complexity of modern AI systems, which may resist simple forms of interpretation. Ethics-by-design thus calls for a nuanced approach to transparency, one that is sensitive to the needs of different audiences and the specificities of different use cases (Mittelstadt, 2019, p. 505).

The principle of human oversight, often operationalized through the notion of "human-in-the-loop" or "human-on-the-loop" systems, represents another cornerstone of ethical integration. While automation can enhance efficiency and consistency, it also risks diminishing human control and responsibility. Embedding ethics into AI therefore requires the design of governance structures that preserve meaningful human involvement in decision-making processes. This does not imply a rejection of automation, but rather a reconfiguration of the relationship between human and machine, in which human judgment remains the ultimate point of normative reference. Effective oversight mechanisms must ensure that human actors are not reduced to passive supervisors, but are empowered to understand, intervene in, and ultimately override algorithmic decisions when necessary.

Beyond these technical and procedural mechanisms, ethics-by-design also necessitates the development of institutional governance models capable of supporting ethical integration at an organizational level. This includes the establishment of interdisciplinary ethics committees, the integration of ethical expertise into development teams, and the creation of internal accountability structures that align organizational incentives with ethical objectives. Importantly, ethical governance cannot be confined to isolated units within an organization; it must be embedded within the broader organizational culture, influencing decision-making at all levels. This cultural dimension is often overlooked, yet it plays a critical role in determining whether ethical principles are genuinely internalized or merely formally acknowledged.

At the systemic level, the implementation of ethics-by-design raises the question of how different governance actors – public authorities, private companies, civil society organizations, and technical communities – can coordinate their efforts. AI systems are rarely developed or deployed within a single institutional context; they are part of complex ecosystems that span multiple sectors and jurisdictions. Effective ethical governance therefore requires mechanisms for multi-level and multi-stakeholder coordination, including regulatory guidance, standard-setting initiatives, certification schemes, and platforms for knowledge exchange. In the European context, such coordination is particularly important given the diversity of national legal systems and institutional arrangements within the Union.

Despite the availability of these mechanisms, the operationalization of ethics-by-design faces a number of significant challenges. One of the most persistent is the translation problem**,** namely the difficulty of converting abstract ethical principles into concrete technical specifications. While concepts such as fairness or accountability are widely endorsed, their precise meaning often remains contested, and different interpretations may lead to different design choices. This plurality is not necessarily a weakness, but it complicates the development of standardized approaches and raises questions about how ethical trade-offs should be resolved in practice (Mittelstadt, 2019, p. 507).

Another major challenge is the risk of formalization without substance. As ethical requirements become institutionalized, there is a danger that they will be reduced to procedural compliance, losing their critical and reflective dimension. Checklists, templates, and standardized metrics can facilitate implementation, but they may also encourage a superficial approach in which ethical considerations are treated as boxes to be ticked rather than as ongoing processes of evaluation and deliberation. Avoiding this outcome requires a balance between formalization and reflexivity, ensuring that ethical governance remains both structured and responsive.

The issue of power asymmetry also plays a crucial role in the effectiveness of ethics-by-design. Large technology companies possess significant resources, technical expertise, and influence over the development of AI systems, which may enable them to shape ethical standards in ways that align with their own interests. At the same time, smaller organizations and public institutions may lack the capacity to implement complex ethical governance mechanisms. Addressing these asymmetries requires regulatory support, capacity-building initiatives, and the development of shared standards that lower barriers to ethical integration.

Finally, the global nature of AI development introduces an additional layer of complexity. While the European Union promotes a human-centric approach grounded in fundamental rights, AI systems are often developed and deployed across multiple jurisdictions with differing normative frameworks. This raises questions about the translatability and scalability of ethics-by-design, as well as the potential for regulatory fragmentation. In this context, the European model may serve as a normative reference point, but its effectiveness will depend on its ability to engage with broader international dynamics.

Taken together, these mechanisms and challenges suggest that ethics-by-design should be understood not as a single tool or method, but as a comprehensive governance paradigm that integrates ethical reasoning into the structures, processes, and cultures of AI development. Its success depends on the alignment of technical design, institutional arrangements, and regulatory frameworks, as well as on the willingness of stakeholders to engage in continuous reflection and adaptation (Veale, 2021, p. 109).

In this perspective, the embedding of ethics into AI systems is not merely a technical task, but a fundamentally normative project that seeks to redefine the relationship between technology and society. By transforming ethical principles into operational norms, ethics-by-design offers a pathway toward a more responsible and human-centered form of innovation – one that does not merely comply with legal requirements, but actively contributes to the realization of shared values. The final section of this article will synthesize these insights and outline the broader implications for the future of AI governance in the European Union.

Conclusion. The analysis developed in this article has demonstrated that the contemporary European approach to artificial intelligence governance, while normatively ambitious and institutionally advanced, remains at a critical transitional stage. The adoption of the EU AI Act marks a significant achievement in establishing a structured and enforceable regulatory framework grounded in the protection of fundamental rights and the promotion of trustworthy AI. Yet, as the preceding sections have argued, the effectiveness of this framework ultimately depends on its ability to move beyond a narrowly defined compliance-oriented logic and toward a deeper integration of ethical principles into the very architecture of technological systems.

At the center of this transformation lies a fundamental conceptual shift: from viewing law as an external mechanism of constraint to understanding it as part of a broader normative ecosystem in which ethics plays a constitutive role. The risk-based regulatory model, despite its practical advantages, is inherently limited in its capacity to address the full spectrum of ethical challenges posed by artificial intelligence. By prioritizing measurable and foreseeable harms, it risks overlooking more diffuse and systemic effects, including the gradual erosion of human autonomy, the reconfiguration of social relations through algorithmic mediation, and the normalization of opaque decision-making

processes. These phenomena cannot be adequately captured through risk categorization alone, as they are embedded in the design choices and operational logics that shape AI systems from their inception (Veale, 2021, p. 111).

It is precisely in this context that the paradigm of ethics-by-design emerges as a necessary complement to existing regulatory approaches. As argued throughout this article, ethics-by-design should not be understood as a voluntary or supplementary practice, but as a structural condition for the legitimacy of AI governance. By embedding ethical principles – such as human dignity, autonomy, fairness, accountability, and transparency – into the lifecycle of AI systems, it becomes possible to bridge the gap between formal legality and substantive justice. In doing so, ethics-by-design redefines the role of ethics from an external evaluative framework to an internal design logic that actively shapes technological development.

The contribution of this study lies in its attempt to reconceptualize ethics-by-design as a form of operational normativity, capable of translating abstract ethical values into concrete mechanisms of governance. Through the analysis of ethical impact assessments, algorithmic auditing, transparency tools, and human oversight structures, the article has shown that ethical integration is not merely a theoretical aspiration, but a practical possibility – albeit one that requires careful institutional design and continuous reflexive engagement. At the same time, the identified challenges – ranging from the translation of ethical principles into technical specifications, to the risk of formalization without substance, and the persistence of power asymmetries – underscore the complexity of this endeavor and the need for sustained interdisciplinary collaboration.

From a broader perspective, the transition beyond compliance reflects a deeper normative ambition within the European project: the aspiration to align technological innovation with the preservation of human-centered values. This ambition is particularly significant in a global context characterized by competing models of digital governance, where the balance between efficiency, control, and individual rights is negotiated in different ways. The European Union's emphasis on human-centric AI represents not only a regulatory strategy, but also a normative statement about the kind of digital society it seeks to construct – one in which technology serves as an instrument of human flourishing rather than a mechanism of domination or exclusion (Morley, 2020, p. 2163).

However, realizing this vision requires more than the articulation of principles or the adoption of legal instruments. It demands a transformation in how artificial intelligence is conceptualized, developed, and governed. In particular, it calls for a reconfiguration of responsibility, in which ethical considerations are distributed across the entire network of actors involved in the AI lifecycle, and where accountability is embedded within the design of systems rather than imposed solely through external enforcement. It also requires a cultural shift within organizations and institutions, fostering an environment in which ethical reflection is not perceived as an obstacle to innovation, but as a prerequisite for its sustainability and legitimacy.

In this light, the relationship between the EU AI Act and ethics-by-design should be understood as mutually reinforcing rather than hierarchical. The Act provides the legal foundation necessary to establish minimum standards and ensure accountability, while ethics-by-design offers the normative orientation required to guide the interpretation and implementation of these standards in practice. Together, they form the basis of a more integrated model of AI governance – one that combines legal certainty with ethical depth, and procedural rigor with substantive legitimacy.

The findings of this article also suggest several broader implications for future research and policy development. First, there is a need to further refine the methodologies and tools through which ethical principles can be operationalized, ensuring that they remain both practically applicable and normatively meaningful. Second, greater attention should be devoted to the institutional conditions that enable or constrain ethical integration, including issues of capacity, expertise, and organizational culture. Third, the global dimension of AI governance must be taken into account, as the effectiveness

of the European model will depend on its ability to interact with and influence international regulatory and ethical frameworks.

Ultimately, the shift toward a human-centric paradigm of AI governance is not merely a technical or regulatory adjustment, but a profound normative project that redefines the relationship between technology, law, and society. The move beyond compliance signifies a recognition that legal regulation, while necessary, is not sufficient to ensure the ethical legitimacy of artificial intelligence. Only by embedding ethical reasoning into the core of technological design can it be ensured that AI systems remain aligned with the values that underpin democratic societies.

In this sense, the future of AI governance in Europe will not be determined solely by the precision of its legal rules, but by its capacity to preserve and institutionalize the primacy of human agency in an increasingly automated world. Ethics-by-design, as conceptualized in this article, offers a pathway toward achieving this goal, transforming artificial intelligence from a source of normative uncertainty into a domain of responsible and value-oriented innovation.

**References:**


1. European Commission. (2019). *Ethics guidelines for trustworthy AI*. High-Level Expert Group on Artificial Intelligence.
2. European Commission. (2020). *Assessment list for trustworthy artificial intelligence (ALTAI)*.
3. European Commission. (2021). *Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*.
4. European Parliament & Council of the European Union. (2024). *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*.
5. Floridi, L., Cowls, J., Beltrametti, M., et al. (2018). AI4People – An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*(4), 689–707.
6. Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review, 1*(1).
7. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399.
8. Mittelstadt, B. D. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence, 1*(11), 501–507.
9. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2).
10. Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and Engineering Ethics, 26*, 2141–2168.
11. OECD. (2019). *OECD principles on artificial intelligence*.
12. Stix, C. (2021). Actionable principles for artificial intelligence policy: Three pathways. *Science and Engineering Ethics, 27*(1).
13. Veale, M., & Borgesius, F. Z. (2021). Demystifying the draft EU Artificial Intelligence Act. *Computer Law Review International, 22*(4), 97–112.
14. Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology, 31*(2), 841–887.