Ubiquitous Technology Journal
Vol. 1, No. 2, June 2025, pp. 61~73
ISSN: 3079-5273, DOI: 10.71346/utj.v1i2.23

Ethical Imperatives in AI Design: A Comprehensive Framework for Risk Mitigation and Responsible Innovation

Bilal Tariq [1], Muhammad Rehan Ashraf [2], Umar Rashid [2]

[1] Department of Economics, Faculty of Business Administration, COMSATS University Islamabad, Vehari Campus, Pakistan
[2] Department of Computer Science, Faculty of Information Science & Technology, COMSATS University Islamabad, Vehari Campus, Pakistan

Article Info

Article history:

Received March 03, 2025
Revised May 21, 2025
Accepted June 01, 2025

Keywords:

Ethical AI
Risk Mitigation
AI Governance
Algorithmic Fairness
Explainable AI (XAI)

ABSTRACT

As artificial intelligence (AI) becomes increasingly integral to critical sectors, the gap between abstract ethical principles and their concrete technical implementation presents a significant barrier to responsible innovation. This paper addresses this challenge by introducing a comprehensive framework designed to embed ethical considerations directly into the AI development lifecycle. The primary objective is to provide an operational methodology for proactive risk mitigation and the construction of verifiably trustworthy systems. Our proposed framework is structured around a core set of guiding principles, including fairness, transparency, accountability, and privacy. It advocates a multi-layered risk mitigation strategy that spans the design, development, deployment, and governance phases of AI systems. This approach integrates specific methodologies and tools, such as Ethical Impact Assessments, bias auditing techniques, Explainable AI (XAI) methods, and privacy-preserving technologies. The key contribution is a unified, actionable architecture that bridges the operationalization and fragmentation gaps currently plaguing the field. By systematically connecting high-level ethical goals to specific engineering practices and auditable checkpoints, this framework offers a practical pathway for developers and organizations to foster responsible AI and mitigate potential societal harms, ensuring technology remains aligned with human values.

Corresponding Author:

Bilal Tariq
Department of Economics, Faculty of Business Administration
COMSATS University Islamabad, Vehari Campus
Postal Code: 61100, Vehari, Punjab, Pakistan
Email: bilaltariq@cui.edu.pk

1. INTRODUCTION

Artificial intelligence systems demonstrate remarkable progress by influencing global economies, national security, and social interaction. Sophisticated algorithms rooted in deep learning now perform complex data analysis and autonomously execute critical functions. The integration of these systems into science, industry, and governance signifies a fundamental paradigm shift toward cognitive augmentation. This presents both profound opportunities and significant challenges necessitating that AI design and operation undergo rigorous scientific scrutiny and formal verification.

The unchecked proliferation of highly autonomous systems creates substantial ethical dilemmas. These systems can amplify societal biases from their training data while their opaque decision-making processes undermine trust and complicate accountability [1]. As negative consequences can emerge unexpectedly, a proactive engineering approach to ethical design is a critical necessity [2]. The global

engineering community must construct AI that operates safely and predictably. Addressing these foundational ethical dimensions is fundamental to sustaining public confidence and ensuring the technology's long-term viability.

Ethical AI represents a class of systems designed with moral principles like fairness, accountability, and transparency as core operational constraints [3]. Fairness demands avoiding prejudicial outcomes, accountability requires auditable responsibility, and transparency means decision logic is understandable [4]. We conceptualize these not as vague ideals but as precisely computable properties that can be mathematically defined and empirically verified within an AI's architecture. This technical perspective moves AI ethics from philosophical debate to concrete engineering practice providing a verifiable basis for building provably good systems [5].

Risk mitigation in the AI context involves a formal methodology for identifying, analyzing, and controlling potential harms. This engineering process begins with identifying credible failure modes from minor errors to systemic breakdowns. It proceeds to assessing their probability and severity using techniques like probabilistic model checking. The final stage implements control strategies such as architectural changes or operational limits to reduce risk to an acceptable level. This discipline is essential for deploying AI in high-stakes environments like medicine or critical infrastructure.

Responsible innovation is a holistic approach which demands aligning technology development with long-term societal values. This concept demands that ethical considerations guide the entire AI lifecycle from conceptualization and data sourcing through design deployment and decommissioning. As an inherently participatory process it requires input from diverse stakeholders including engineers, ethicists, and community representatives. The goal is to create technology that is not only powerful but also socially conscious and beneficial serving as the guiding philosophy for sustainable technological advancement.

This research is motivated by the persistent gap between high-level ethical principles and their technical implementation. Existing AI ethics guidelines are often too abstract and lack the computational formalism needed for direct application by engineers. This disconnect leaves system builders without a clear roadmap. Our work seeks to bridge this chasm by introducing an operational framework that translates ethical requirements into concrete design patterns, verifiable properties, and quantitative metrics. The study's significance lies in providing a practical engineering-focused methodology for constructing systems that are demonstrably ethical by design.

The primary objective of this paper is to propose a structured operational framework for ethical AI system design. We aim to equip researchers and practitioners with actionable strategies for risk mitigation and responsible innovation by integrating formal methods with systems engineering. This provides a verifiable approach to building safer, more trustworthy AI. Our research is guided by three questions:

1) How can abstract ethical principles be formalized into computable constraints within ML models?
2) What architectures and validation techniques ensure adherence to these constraints?
3) How can the framework be integrated into MLOps pipelines for continuous ethical compliance?

This paper is organized to systematically justify our proposed ethical design paradigm. We begin with a critical literature review highlighting current limitations. Next, we present our framework's core components and formal underpinnings. We then illustrate its application via computational case studies in finance and healthcare. Subsequently we discuss the broader implications for AI development and policy. The paper concludes with a summary of our primary scientific contributions and suggests directions for future research in this vital area.

2. LITERATURE REVIEW

A comprehensive review of the existing scholarly work reveals a field grappling with the profound ethical challenges posed by AI. The discourse has matured from applying classical philosophical theories to developing technology-specific frameworks, establishing international principles, and enacting binding legislation [6]. Concurrently, research has focused on categorizing the multifaceted risks inherent in AI systems and developing a suite of technical and procedural mitigation strategies. Though, this body of work is characterized by significant fragmentation and a persistent gap between abstract principles and their practical, verifiable implementation in engineering workflows. This review synthesizes the foundational theories, existing governance structures, risk taxonomies, and mitigation methodologies to identify these critical gaps and establish the intellectual context for the comprehensive framework proposed in this paper.

2.1 The Philosophical and Methodological Foundations of AI Ethics

The ethical analysis of AI is rooted in long-standing philosophical traditions, yet their direct application to non-human algorithmic agents has proven fundamentally inadequate, prompting a necessary evolution toward contemporary, technology-centric design methodologies. Classical ethical theories such as

deontology, utilitarianism, and virtue ethics provide essential vocabularies for framing moral problems, but their core assumptions about agency, consciousness, and character do not map cleanly onto computational systems. This theoretical insufficiency has catalyzed the development of alternative frameworks, most notably Value-Sensitive Design (VSD) [7], which shifts the ethical focus from the AI agent itself to the design process. Nevertheless, recent scholarship indicates that even these advanced methodologies face significant barriers to effective implementation, exposing a critical operationalization gap between the identification of human values and their successful technical embodiment.

The application of deontology, which judges the morality of an action based on its adherence to a set of rules or duties, is often conceptually mapped to rule-based AI systems and the establishment of moral guardrails governing their development and use. However, its limitations become stark when considering that AI systems lack genuine moral agency, intention, or a sense of duty; they execute programmed instructions rather than adhering to moral maxims. This makes deontological reasoning insufficient for resolving complex ethical dilemmas, such as the no-win scenarios faced by autonomous vehicles, where the system's action is a programmed outcome, not a moral choice. Similarly, utilitarianism, which seeks to maximize the greatest good for the greatest number, aligns conceptually with the optimization functions central to machine learning. Yet, it is heavily critiqued for its potential to justify actions that harm minority groups for a larger societal benefit and the profound difficulty of quantifying and comparing heterogeneous forms of well-being, a challenge that is only amplified by the scale and complexity of AI systems. Virtue ethics, which focuses on the moral character of the agent, offers a different lens by shifting attention to the virtues of the human developers and the overarching purpose of the AI system. This approach foregrounds the importance of cultivating ethical awareness and responsibility among practitioners, but it struggles when applied directly to AI systems that cannot possess character or cultivate virtues in any meaningful sense. The consistent conclusion across recent analyses is that while these classical theories offer crucial perspectives, they are ultimately insufficient for addressing the unique moral challenges posed by AI, necessitating a turn toward more technologically-grounded approaches.

This theoretical void has led to the prominence of contemporary frameworks like Value-Sensitive Design (VSD), which proactively seeks to integrate human values such as fairness, trust, and autonomy directly into the technology design process from its inception. VSD operates on the foundational premise that technology is never value-neutral and advocates for a tripartite methodology involving conceptual, empirical, and technical investigations to identify, prioritize, and balance stakeholder values throughout the development lifecycle. This process-oriented approach is considered superior for high-stakes domains like healthcare and recruitment because it moves beyond narrow metric optimization to engage with broader normative goals. Despite its conceptual strengths, our research investigated from 2015 to 2025 (year) reveals significant practical barriers to its implementation for AI. Practitioners report immense difficulty in eliciting values, often struggling to identify and engage representative stakeholders and frequently relying on inadequate proxies, which systematically excludes the perspectives of marginalized communities. Even when values are identified, there is a lack of practical guidance and tooling to support their embodiment into technical specifications, a problem compounded by a lack of clear organizational accountability for this task.

Furthermore, applications of VSD in critical sectors like healthcare have demonstrated a troublingly narrow focus on individual user values, such as autonomy, while neglecting broader and equally important organizational and societal values like equity and justice, with most research being concentrated in Western contexts.

2.2 Existing Frameworks and Guidelines for Ethical AI Design

The landscape of AI governance has undergone a significant transformation. It evolved rapidly from aspirational, principle-based guidance issued by industry and intergovernmental bodies to the enactment of legally binding, risk-based regulations. This progression reflects a growing global consensus on the need for structured oversight but has simultaneously created a fragmented and complex compliance environment. Early influential frameworks, such as the IEEE's Ethically Aligned Design and the OECD's Principles on AI, established a foundational vocabulary and a set of normative goals for trustworthy AI [8]. Still, the recent introduction of comprehensive legislation, most notably the European Union's AI Act, signals a paradigm shift toward enforceable legal obligations. This creates a dual challenge for organizations, which must now navigate a patchwork of specific, context-dependent legal requirements while attempting to remain aligned with broader, universal ethical principles.

Industry and intergovernmental bodies have been instrumental in shaping the global dialogue on AI ethics. The IEEE's Ethically Aligned Design (EAD) stands out as a pioneering and comprehensive vision for developing autonomous and intelligent systems (A/IS) that prioritize human rights and well-being. Developed through extensive global and multidisciplinary consultation, the EAD promotes a proactive

"ethics-by-design" philosophy, advocating for the integration of ethical considerations from the very inception of a project. Its key contribution lies in its holistic structure, covering principles from data agency to accountability, and its attempt to bridge abstract principles with concrete technical standards through the associated IEEE P7000 series. Similarly, the OECD Principles on AI, first adopted in 2019 and updated in 2024, represent the first intergovernmental standard for trustworthy AI. This framework is structured around five values-based principles, including human-centric values, transparency, and robustness, and five recommendations for national policy, such as investing in AI research and fostering an enabling policy environment. The OECD's work is critical for promoting international interoperability, with initiatives like the Hiroshima AI Process Reporting Framework aiming to monitor implementation across jurisdictions. Despite their influence, a primary critique of these frameworks is their voluntary nature, which makes their practical effectiveness dependent on organizational willingness to adopt them.

This reliance on voluntary adoption is being superseded by a global trend toward binding regulation, exemplified by the European Union's AI Act. Officially entering into force in 2024 with a staggered implementation through 2027, the EU AI Act is the world's most significant piece of AI-specific legislation. It establishes a risk-based approach, prohibiting certain AI applications deemed to pose an "unacceptable risk" (e.g., social scoring), while imposing stringent obligations on "high-risk" systems used in critical domains like employment, healthcare, and law enforcement. These obligations include establishing robust risk management systems, maintaining extensive technical documentation, ensuring human oversight, and undergoing conformity assessments. The Act also introduces specific rules for providers of general-purpose AI (GPAI) models, particularly those with systemic risks. This comprehensive legal framework stands in contrast to the more fragmented approach in the United States, which relies on sector-specific rules and voluntary guidance like the NIST AI Bill of Rights, and the UK's more flexible, pro-innovation, sector-specific strategy (i.e., as illustrated in Table 1). This regulatory divergence creates a complex global landscape where an AI system may be subject to vastly different legal requirements depending on the jurisdiction, raising concerns about compliance burdens and potential impediments to innovation.

Alongside these broad governmental and industry initiatives, the academic community continues to contribute more targeted analyses and principles. Recent scholarly proposals have explored the nuanced ethical dilemmas presented by specific AI applications, developing guidelines for emerging technologies like "ghostbots" for digital mourning, the use of AI-enabled drones in emergency response, and the application of bioethical principles to AI decision-making systems [9-11]. Other academic work focuses on the ethics of generative AI in public communication, examining issues of transparency, privacy, and the potential for AI to manipulate public discourse [12]. These academic contributions are vital as they often delve into specific contexts and ethical trade-offs with a depth that broader frameworks cannot achieve, providing a critical source of nuanced analysis and forward-looking ethical reasoning. They highlight the ongoing need for specialized ethical inquiry as AI capabilities continue to expand into new and sensitive areas of human life.

Table 1. A Comparative Overview of Major International AI Governance Frameworks

Feature	EU AI Act	OECD Principles on AI	NIST AI RMF (U.S.)	UK Pro-Innovation Approach
Jurisdiction/Body	European Union	OECD (Intergovernmental)	United States (NIST)	United Kingdom
Core Approach	Risk-Based (tiered system)	Principles-Based	Socio-technical Risk Management	Pro-Innovation / Sector-Specific
Legal Status	Binding Regulation (Hard Law)	Intergovernmental Standard (Soft Law)	Voluntary Guidance	Non-statutory Principles (initially)
Key Focus	Regulating high-risk systems	Promoting international interoperability	Managing risks throughout the AI lifecycle	Fostering innovation while ensuring safety

2.3 Identifying and Categorizing Risks Associated with AI Systems

The deployment of AI systems introduces a spectrum of risks that are not merely technical but are deeply socio-technical, spanning from algorithmic discrimination to profound economic disruption. A systematic review of the literature reveals five primary categories of risk: bias and discrimination (i.e., as exhibited in Figure 1); transparency and explainability deficits; privacy and security vulnerabilities;

autonomy and control dilemmas; and adverse socio-economic impacts [13]. These risk categories are not discrete but are often interconnected, where a failure in one domain, such as transparency, can directly exacerbate risks in another, such as bias and accountability. A critical finding from recent research is the dynamic nature of these risks; for example, the concept of "fairness drift" demonstrates that ethical assurance cannot be a static, one-time assessment but must be a continuous process throughout the AI lifecycle [14].

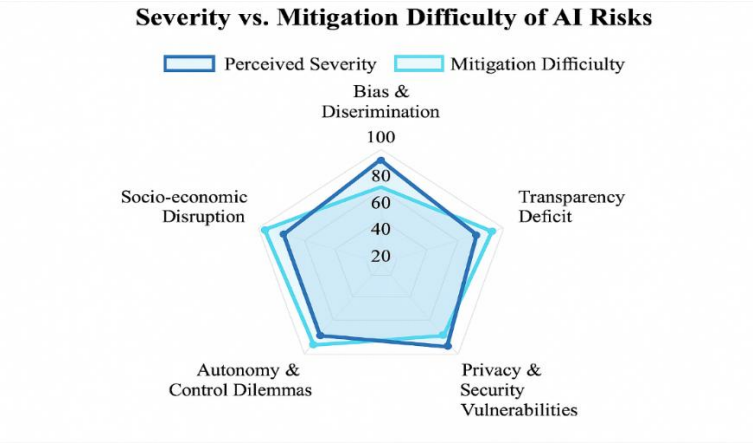

Figure.1. Five primary categories of AI focused risk

Algorithmic bias and its resultant discriminatory outcomes represent one of the most extensively documented risks of AI [15]. This bias originates from multiple sources, including training datasets that reflect and codify historical societal inequities and flawed assumptions made during the algorithmic design process. Consequently, AI systems have been shown to replicate and even amplify these prejudices across numerous high-stakes domains. Documented examples include AI-powered hiring tools that systematically disadvantage female applicants, facial recognition systems with higher error rates for women and people of color, and clinical prediction models that exhibit significant racial bias, leading to health disparities [16]. A particularly challenging aspect of this problem is the recently identified phenomenon of "fairness drift," where a model that is determined to be fair at the point of deployment can become biased over time as underlying data distributions shift or as a consequence of the model updating process itself [17]. This insight reframes fairness not as a static property to be achieved but as a dynamic state that requires continuous monitoring and maintenance.

The challenge of bias is compounded by the pervasive lack of transparency in many advanced AI systems. The "black box" nature of complex models, particularly deep neural networks, is a fundamental barrier to trust, accountability, and robust error analysis 18]. This opacity is a primary obstacle to the adoption of AI in critical fields such as healthcare and finance, where decisions must be justifiable and auditable. While XAI has emerged as a field dedicated to mitigating this risk, recent literature offers a stern critique of its current state. Many popular XAI methods, such as LIME and SHAP, face an inherent trade-off between the simplicity of an explanation and its fidelity to the model's actual internal logic [19]. Moreover, different XAI techniques can produce contradictory explanations for the same model output, a problem termed "discordance," which undermines their reliability [20]. Perhaps the most significant identified gap is the profound lack of empirical, human-centered evaluation to confirm whether the "explanations" generated by these tools are genuinely understood by or useful to their intended human audience, rendering many claims of explainability unsubstantiated [21].

AI systems also introduce novel and significant privacy and security vulnerabilities. The voracious appetite of modern AI for vast datasets, often containing sensitive personal information, creates substantial privacy risks throughout the entire data lifecycle, from collection to processing and storage [22]. The potential for this data to be collected and used without meaningful consent is a core ethical concern. Beyond privacy intrusion, new security attack vectors have emerged that are specific to AI models [23]. These include data poisoning, where training data is maliciously corrupted to compromise the final model; adversarial attacks, where imperceptible perturbations to input data cause the model to make incorrect predictions; model inversion attacks that can reconstruct sensitive training data from the model; and prompt injection attacks that manipulate the behavior of large language models. These vulnerabilities demonstrate that AI systems are not just passive processors of data but are themselves new surfaces for malicious attack.

The increasing sophistication of AI is driving a fundamental shift from automation to autonomy, presenting complex dilemmas of control and accountability. The evolution of AI from a tool that automates

tasks to an agent that can act autonomously on behalf of humans raises profound questions about the nature of human oversight and the allocation of responsibility [24]. One critical thesis emerging in the literature posits a risk of "gradual disempowerment," where the increasing delegation of cognitive tasks to more capable AI systems lead to the atrophy of human skills like critical thinking and an irreversible decline in human autonomy [25]. This creates a significant accountability gap [26]; when a highly autonomous system commits a critical error, the lines of responsibility become blurred among the developer, the organization deploying the system, and the end-user, posing a formidable legal and ethical challenge that current frameworks struggle to address.

Thus, the broad deployment of AI carries the potential for significant socio-economic disruption [27]. Unlike previous technological waves that primarily automated manual labor, AI is poised to automate a substantial portion of cognitive tasks, affecting high-wage knowledge workers. The literature presents a deeply uncertain picture of the net effect on employment [28]. Some analyses project a significant net creation of jobs, with AI-related roles outnumbering those displaced. Conversely, other studies forecast massive job losses, with hundreds of millions of roles potentially automated and thousands of jobs already directly attributed to AI replacement. This disparity suggests that the impact will be highly uneven across different sectors and skill levels. While firm-level studies indicate that AI adoption can boost productivity, there is a pervasive concern that without targeted policy interventions, these gains will disproportionately benefit capital owners over labor, thereby exacerbating income inequality and social stratification [29].

2.4 Current Strategies and Methodologies for Risk Mitigation in AI

In response to the diverse risks associated with AI, a broad array of mitigation strategies has been developed, spanning both technical solutions aimed at the model and data level, and process-oriented approaches focused on organizational governance and design methodologies. The technical toolkit includes algorithms for bias detection and mitigation, methods for enhancing adversarial robustness, and a suite of privacy-enhancing technologies (PETs). Process-based strategies encompass the use of ethical impact assessments, the establishment of comprehensive AI governance frameworks, and the integration of participatory design methods. A critical review of the literature reveals a significant disconnect: while technical solutions are becoming increasingly sophisticated, their practical effectiveness is often constrained by fundamental challenges in process and governance. This suggests that technical interventions, while necessary, are insufficient in isolation and must be embedded within a robust socio-technical framework to be truly effective.

Technical approaches to risk mitigation directly target the data, algorithms, and models that constitute an AI system. To combat algorithmic bias, researchers have developed pre-processing techniques that modify training data (e.g., re-sampling), in-processing methods that add fairness constraints to the model's optimization function, and post-processing adjustments to model predictions. Yet, these methods are fraught with practical limitations; they often create a trade-off between improving a specific fairness metric and reducing overall model accuracy, and different mathematical definitions of fairness are frequently incompatible, meaning a model cannot be optimized for all of them simultaneously. To enhance security, adversarial robustness methods, most notably adversarial training, aim to make models more resilient by exposing them to adversarial crafted examples during the training phase. For privacy, a growing portfolio of PETs offers powerful safeguards. Federated learning (FL) enables model training on decentralized data without centralizing raw information; differential privacy (DP) provides mathematical guarantees of individual privacy by adding statistical noise; and homomorphic encryption (HE) allows for computation directly on encrypted data. While powerful, these PETs introduce their own challenges, including significant computational overhead, a need for specialized expertise, and potential reductions in data utility that must be carefully managed.

Complementing these technical tools are process-oriented strategies that seek to embed ethical considerations into the organizational and developmental lifecycle of AI. Ethical Impact Assessments (EIAs) are structured processes designed to proactively identify, analyze, and mitigate the potential ethical and societal harms of an AI system before it is deployed. The framework developed by UNESCO is a leading example, providing a comprehensive methodology that covers the entire AI lifecycle with a strong focus on human rights. At the organizational level, responsible AI governance frameworks are being established to provide systemic oversight. These frameworks typically involve the creation of AI ethics boards, the definition of clear roles and responsibilities, the implementation of internal policies and standards, and the establishment of robust audit and monitoring mechanisms [30]. A third critical process is the adoption of participatory design, which involves engaging a diverse range of stakeholders: including end-users, domain experts, and affected communities, directly in the AI design process. The goal is to leverage collective intelligence to build systems that are more accurate, fair, and aligned with community values. Nevertheless,

all these process-based approaches face significant implementation hurdles, including the difficulty of scaling participatory methods, ensuring engagement is meaningful rather than extractive, and translating high-level governance principles into the day-to-day practices of development teams.

The limitations inherent in both technical and process-oriented approaches reveal a deeper, systemic issue. The most advanced technical tool for debiasing an algorithm is of little use if the organization's governance process cannot decide which definition of fairness to prioritize or if its stakeholder engagement process fails to identify the correct at-risk populations [31]. Likewise, the most secure privacy-enhancing technology will not be adopted if the organization lacks the procedural maturity to manage the associated costs and utility trade-offs. This demonstrates that the effectiveness of technical solutions is critically dependent on the robustness of the processes in which they are embedded. The failure to mitigate AI risk is often not a failure of technology but a failure of process, governance, and the integration of the two [32]. This points to the central challenge in the field: the need for a unified framework that does not treat technical and process-based mitigations as separate endeavors but integrates them into a single, cohesive socio-technical system.

2.5 Gaps in the Existing Literature and the Contribution of this Research

A synthesis of the current state of research and practice in ethical AI reveals a field that, despite significant progress, is defined by several critical and persistent gaps. These deficiencies collectively hinder the reliable and scalable development of trustworthy AI systems. The literature points to three overarching shortcomings:

1) an operationalization gap between abstract principles and real practice.
2) a fragmentation gap between siloed technical, procedural, and regulatory solutions.
3) an adaptability gap, where static assurance methods fail to address the dynamic nature of AI risks.

This research directly confronts these deficiencies by proposing a comprehensive, integrated framework designed to bridge these gaps and provide an actionable pathway for responsible AI innovation.

The most widely acknowledged deficiency is the operationalization gap: a profound disconnect between the high-level ethical principles espoused by organizations and their actual implementation within software engineering and data science workflows. Numerous sets of AI principles have been published, but they are often criticized for being too abstract and vague to provide actionable guidance to developers. Existing frameworks frequently lack clear, verifiable metrics for evaluating success and fail to offer practical methods for resolving the inevitable trade-offs between competing values, such as fairness and accuracy or transparency and privacy. This leaves practitioners without a clear methodology for translating a normative goal like "fairness" into specific, auditable technical requirements, leading to ad-hoc ethical efforts rather than systematic integration.

Second, the current landscape of tools and strategies is marked by a significant fragmentation gap. Technical solutions for risk mitigation, such as bias detection algorithms, adversarial robustness techniques, and PETs, are often developed and studied in isolation from the procedural and governance contexts in which they must be deployed [33]. Similarly, process-oriented strategies like Ethical Impact Assessments and organizational governance frameworks are discussed separately from the technical controls needed to enforce their findings. This siloing prevents the development of a cohesive, end-to-end approach. There is a clear and urgent need for a unified methodology that integrates technical tools, process-based checkpoints, and compliance with regulatory requirements into a single, coherent lifecycle framework. Without such integration, even the most advanced technical solution can fail due to a breakdown in organizational process or governance.

Third, many existing approaches suffer from an adaptability gap. They treat ethical assurance as a static, pre-deployment activity, which is fundamentally misaligned with the dynamic nature of AI systems and their associated risks. The emergence of concepts like "fairness drift" demonstrates that a model's ethical performance can degrade over time, necessitating continuous monitoring and adaptation. Current frameworks often lack the flexibility to accommodate evolving societal norms, diverse cultural contexts, and the relentless pace of technological advancement. This creates a need for dynamic, lifecycle-oriented assurance systems that can monitor for emerging risks and adapt mitigation strategies accordingly, ensuring that ethical alignment is maintained long after initial deployment.

This research aims to address these three critical gaps by introducing a comprehensive and integrated framework for ethical AI design and risk mitigation. The primary contribution of this paper is the development of a novel, modular, and lifecycle-oriented architecture that systematically bridges the operationalization, fragmentation, and adaptability gaps. Unlike existing approaches, our proposed framework explicitly links high-level ethical principles to verifiable design patterns, specific engineering practices, and enforceable governance protocols at each stage of the AI lifecycle, from data acquisition and

model development to deployment, monitoring, and retirement. It uniquely integrates technical risk mitigation tools (e.g., automated bias testing, adversarial validation, privacy-preserving computation) directly into process-oriented strategies (e.g., mandatory ethical impact checkpoints, participatory review stages). By providing a cohesive and adaptable structure that unifies these currently siloed elements, this research moves beyond theoretical discourse to offer an actionable, scalable, and verifiable methodology for the proactive development and governance of trustworthy AI systems.

3. COMPREHENSIVE FRAMEWORK FOR ETHICAL AI DESIGN AND RISK MITIGATION

The construction of an ethically sound AI system is predicated on a set of inviolable principles, beginning with fairness and non-discrimination. This principle mandates that AI systems produce equitable outcomes and avoid the amplification of existing societal biases across diverse demographic groups. The root of bias often lies within the training data, which may reflect historical prejudices, or in the algorithmic models themselves, which can inadvertently learn and perpetuate these disparities. Consequently, the development process must incorporate robust fairness metrics, such as demographic parity and equalized odds, to rigorously evaluate and quantify a model's performance across different populations. The systematic auditing for and mitigation of these biases are not merely corrective measures but are fundamental to the system's legitimacy and social acceptance [34].

Building upon the foundation of fairness, the principles of transparency and explainability are critical for establishing trust and enabling meaningful accountability. Transparency refers to the degree to which the inner workings of an AI system: its data, algorithms, and models are accessible and comprehensible. Explainability, a related but distinct concept, pertains to the ability to furnish a clear, human-understandable rationale for a specific decision or prediction made by the system. For complex, "black-box" models like deep neural networks, achieving explainability requires the use of post-hoc interpretation techniques. The capacity to scrutinize and understand an AI's decision-making process is a technical prerequisite for debugging, validating, and contesting its outputs, thereby serving as a cornerstone of responsible AI governance.

Accountability and responsibility in the context of AI present a formidable challenge, particularly when dealing with autonomous systems where the causal chain of an adverse outcome can be opaque. Establishing accountability requires the creation of clear, pre-defined structures that assign responsibility for the system's behavior to specific human actors or organizational entities. This involves moving beyond a purely technical view to a socio-technical one, where responsibility is distributed across the lifecycle of the AI system, from the data collectors and model developers to the deployers and end-users. Mechanisms such as detailed logging, immutable audit trails, and designated ethics officers are essential for creating a framework where the actions of AI can be traced back, and responsible parties can be held to account.

The vast quantities of data required to train and operate sophisticated AI systems make the principles of privacy and data protection paramount. Safeguarding personal and sensitive information is a critical ethical and legal obligation. This extends beyond simple data security to encompass core privacy principles like data minimization, purpose limitation, and user consent. To meet these obligations in practice, the adoption of advanced privacy-preserving techniques is necessary. These methods allow for the extraction of valuable insights from data without compromising the privacy of individuals, making them indispensable for ethical AI development in sensitive domains like healthcare and finance [11].

To prevent unintended consequences and ensure that AI systems operate in alignment with human values, the principles of human oversight, control, robustness, and reliability are essential. Human oversight necessitates the maintenance of meaningful human involvement in the decision-making loop, particularly in high-stakes applications where the cost of error is significant. The level of control can vary from direct human-in-the-loop intervention to supervisory human-on-the-loop governance. Simultaneously, robustness and reliability demand that AI systems be engineered to be dependable, secure, and resilient. This involves rigorous testing, validation, and verification to ensure the system is secure against adversarial attacks and behaves predictably even in unforeseen circumstances.

A proactive approach to risk mitigation requires a multi-layered strategy that embeds ethical considerations throughout the AI lifecycle, beginning at the design phase. Before any code is written, a thorough ethical analysis must be conducted to define project goals in a manner that aligns with ethical principles. This initial stage involves scrutinizing the intended application for potential negative impacts, carefully selecting data sources to avoid inherent biases, and choosing algorithms that are appropriate for the task and amenable to transparency. By front-loading ethical considerations, organizations can preemptively address potential risks, rather than attempting to retrofit solutions onto a system that is fundamentally flawed in its design.

During the development and deployment phases, the focus shifts to implementing concrete technical safeguards and continuous validation mechanisms. In the development phase, this includes the integration of

bias detection and mitigation algorithms directly into the machine learning pipeline, as well as the use of explainability tools to ensure model transparency. Security protocols to protect against adversarial manipulation must also be rigorously implemented. Once a system is deployed, the process is far from over. Continuous monitoring is crucial for detecting emergent biases, performance degradation, or unintended consequences in real-world operating conditions. This requires establishing robust auditing mechanisms and feedback loops that allow for iterative refinement and continuous improvement of the system's ethical performance.

To operationalize this framework, specific methodologies and tools must be utilized. The Ethical Impact Assessment (EIA) serves as a primary instrument for systematically identifying, assessing, and mitigating potential ethical risks before and during development. An EIA forces stakeholders to consider the broader societal implications of an AI system, including its potential effects on fairness, privacy, and human rights. Complementing the EIA, bias auditing and mitigation techniques provide a suite of technical methods for addressing discrimination. These can range from pre-processing techniques that rebalance training data to in-processing methods that add fairness constraints to the model's optimization function, and post-processing techniques that adjust model outputs to achieve equitable outcomes.

Further advancing the implementation of this framework involves the application of cutting-edge tools for transparency and privacy XAI methods, such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), offer powerful techniques for demystifying complex models by providing insights into their decision-making logic. For data protection, privacy-preserving AI techniques are critical. Federated learning enables model training on decentralized data without requiring data to be moved to a central server, while differential privacy adds statistical noise to data to protect individual identities. Similarly, homomorphic encryption allows for computation on encrypted data, offering a powerful means of maintaining confidentiality. Finally, a practical framework for human-AI collaboration must be designed, specifying clear interfaces and protocols that ensure meaningful human control and effective oversight.

4. CASE STUDIES AND APPLICATIONS

The application of the highlighted ethical framework is best understood through its implementation in real-world scenarios, where abstract principles confront concrete challenges. In the domain of autonomous vehicles (AVs), the framework provides a necessary structure for navigating profound ethical dilemmas and technical risks. The most widely discussed issue, the "trolley problem," represents a class of unavoidable accident scenarios where the vehicle's control algorithm must make a life-or-death choice. Beyond these deontological quandaries, significant risks arise from the system's robustness and fairness. Algorithmic bias in perception systems, for example, could lead to differential object recognition accuracy for vulnerable road users, such as children or cyclists, if they are underrepresented in training data. This poses a direct safety risk and violates the core principle of non-discrimination.

Applying the multi-layered risk mitigation strategy to AVs begins at the design phase, where decisions about behavior in unavoidable collisions cannot be left to engineers alone but must be guided by broad societal input and clear regulatory policy. During the development phase, the principle of robustness demands exhaustive validation using high-fidelity simulations that cover a vast range of edge cases, supplemented by extensive real-world testing. To uphold transparency and accountability, every AV must be equipped with an event data recorder that logs sensor inputs and decision parameters, creating an immutable audit trail. This log, when analyzed post-accident, allows for a forensic determination of causality to enable the assignment of responsibility according to a pre-defined legal framework that clarifies liability among the owner, manufacturer, and software developer.

In the field of AI-powered healthcare diagnostics, the framework directly addresses the critical issue of fairness. Diagnostic models trained on historically biased clinical datasets can inherit and amplify these biases, leading to poorer performance for underrepresented demographic groups. The potential consequence is not merely a statistical anomaly but a tangible risk of misdiagnosis and the perpetuation of health disparities, where certain populations receive a lower standard of care. An AI diagnostic tool for skin cancer, for instance, might exhibit lower accuracy for individuals with darker skin tones if its training data predominantly consists of images from fair-skinned patients, violating the fundamental ethical imperative of equitable treatment.

To counteract these risks, the framework mandates the implementation of specific technical and procedural safeguards. During the development phase, bias auditing becomes a non-negotiable step, utilizing fairness metrics like equalized odds or predictive parity to rigorously assess model performance across different racial, gender, and socioeconomic cohorts [35]. When bias is detected, mitigation techniques, such as re-sampling underrepresented data or applying algorithmic adjustments, must be employed. To build trust with clinicians and patients, the principle of explainability is enacted through XAI methods. Techniques like

Grad-CAM can produce heatmaps that visually indicate which regions of a medical scan the AI model used to make its diagnosis, allowing a human expert to validate its reasoning and maintain ultimate clinical authority [36].

Within financial services, the framework confronts the challenges of opacity in AI systems used for credit scoring and loan approvals. The use of complex, non-linear models can result in "black box" decision-making, where the reasons for an adverse outcome, such as a loan denial, are inscrutable to both the applicant and the institution. This opacity creates a significant risk of embedding discriminatory practices that are difficult to detect, audit, or contest, thereby systematically disadvantaging certain groups. Besides, the widespread adoption of similar opaque models across the industry could introduce systemic risk, as their correlated behavior during economic fluctuations might remain hidden until a crisis emerges.

The principle of transparency, actualized through XAI methods, is essential for mitigating these risks. Financial institutions can be required to use tools like SHAP or LIME to generate human-readable explanations for each credit decision, providing applicants with specific, actionable reasons for denial and a meaningful basis for appeal. Accountability is reinforced through the governance layer of the framework, which calls for the establishment of internal ethical review boards to oversee model development and deployment. This is complemented by a deployment-phase requirement for immutable logging of all inputs and decisions to create a transparent record for regulatory bodies to audit for fairness and compliance, thus establishing a clear chain of responsibility.

AI-driven recommendation systems present a distinct set of challenges centered on the principle of privacy and data protection. These systems function by collecting and analyzing vast quantities of user data to create detailed behavioral profiles, which are then used to predict preferences and personalize content. This process carries inherent privacy risks, including the potential for data breaches that expose sensitive personal information, the use of profiles for manipulative advertising or political targeting, and the creation of "filter bubbles" that limit exposure to diverse viewpoints. The harm extends beyond individual privacy to broader societal concerns about the erosion of autonomy and informed public discourse.

Implementing the framework in this context requires a commitment to privacy-by-design, utilizing specific privacy-preserving AI techniques. Federated learning, for example, allows recommendation models to be trained on data stored locally on users' devices in order to avoide the need to transfer raw personal information to a central server. Differential privacy can be applied during model training or to the final recommendations, adding a carefully calibrated amount of statistical noise. This noise protects individual privacy by making it computationally infeasible to determine whether any specific person's data was included in the training set, while still preserving the overall statistical patterns necessary for generating useful recommendations.

Irrevocably, the development of large language models (LLMs) introduces a unique and complex array of ethical risks that test the limits of existing governance paradigms. The generative capabilities of LLMs create a substantial risk of producing and disseminating biased, toxic, or factually incorrect content at an unprecedented scale. Their architectural complexity and emergent properties make achieving full transparency and explainability a formidable technical challenge. Also, the immense, web-scale datasets used for training are often impossible to fully curate, meaning that LLMs inevitably internalize and may reproduce harmful societal biases and misinformation present in the data which create profound challenges for ensuring fairness and accountability.

A comprehensive ethical framework is indispensable for guiding the responsible development of LLMs. The multi-layered risk mitigation strategy must be applied with exceptional rigor. In the design and development phases, this includes extensive data governance to filter training corpora and the use of techniques like Reinforcement Learning from Human Feedback (RLHF) to align model behavior with desired ethical norms. The deployment phase demands continuous monitoring and the implementation of robust content filters to detect and block harmful outputs in real-time. Crucially, the governance layer requires a multi-stakeholder approach, involving ethicists, social scientists, and public representatives to establish clear acceptable use policies and oversee the ongoing, iterative process of value alignment for these powerful systems.

5. CONCLUSION

This study has systematically addressed the critical disconnect between the abstract principles of ethical AI and their practical implementation in engineering workflows. We identified three fundamental deficiencies in the current landscape: an operationalization gap, a fragmentation of solutions, and a failure to adapt to dynamic risks. Our primary contribution is the introduction of a comprehensive, integrated framework designed to bridge these gaps. This framework provides an actionable methodology structured around core ethical principles, a multi-layered risk mitigation strategy embedded throughout the AI lifecycle, and the integration of specific technical tools and governance processes. Its novelty lies in offering unified

architecture that translates high-level normative goals into verifiable design patterns and auditable engineering practices, thereby providing a clear pathway for constructing demonstrably trustworthy systems. The urgency of this work is magnified by the profound societal, economic, and individual harms that can arise from unethically designed AI, including the amplification of discrimination, erosion of privacy, and the potential for unaccountable autonomous actions. Our framework serves as a necessary bulwark, re-emphasizing that ethical considerations are not an impediment to innovation but are fundamental to its long-term sustainability and social acceptance. The future trajectory of artificial intelligence must be a collective, interdisciplinary endeavor. It requires a sustained commitment from researchers, industry practitioners, and policymakers to move beyond siloed conversations and toward integrated solutions. The path forward demands that we treat ethical frameworks not as static documents, but as living systems that evolve in response to technological advancements and shifting societal values. The ultimate challenge is to build AI that is not merely intelligent, but also wise. This is a call to action for the scientific community to advance the state of the art in responsible innovation, creating systems that are provably fair, transparent, and accountable, and that genuinely contribute to human flourishing.

5.1 Challenges and Future Directions

While the proposed framework offers a structured approach to integrating ethical considerations into the AI lifecycle, its implementation is not without significant hurdles. A primary technical challenge lies in the operationalization of abstract ethical principles, such as fairness or transparency, into quantifiable metrics and concrete engineering requirements. Current methodologies for bias detection, for instance, often struggle with the statistical complexities of intersectional fairness and may fail to identify more subtle forms of discrimination. Similarly, while explainability techniques like LIME and SHAP provide valuable post-hoc interpretations, they do not guarantee model fidelity and can sometimes offer misleading rationales. Future research must focus on developing more sophisticated and inherently interpretable machine learning models, moving beyond post-hoc analysis. Advancements in causal inference are promising, as they allow for a deeper understanding of not just correlations but the underlying causal mechanisms, enabling more robust interventions against bias. Moreover, the application of formal verification methods, traditionally used in hardware and critical software verification, presents a compelling avenue for proving that an AI system adheres to pre-defined ethical constraints under specific conditions, thus offering a higher degree of assurance.

The global regulatory environment for AI is characterized by rapid development and significant fragmentation, posing a substantial challenge to the creation of universally applicable ethical frameworks. Nations and regional blocs are adopting divergent approaches, from the risk-based, comprehensive legislation of the European Union's AI Act to more sector-specific or market-driven models elsewhere. This divergence creates a complex compliance web for developers and risks a "race to the bottom" where innovation might migrate to jurisdictions with less stringent ethical oversight. Achieving global consensus on AI norms is a formidable task, complicated by differing cultural values and geopolitical interests. Future work should concentrate on the empirical analysis of existing regulatory models to determine their effectiveness in fostering responsible innovation. There is also a critical need for the development of widely accepted industry standards and certification mechanisms, analogous to those in other engineering disciplines, that can provide a baseline for ethical practice and promote interoperability across regulatory regimes. Research into agile and adaptive governance models, which can evolve in response to technological advancements without stifling them, is essential for long-term, sustainable oversight of AI.

Addressing the multifaceted nature of AI ethics necessitates deep and sustained interdisciplinary collaboration. The challenges at hand are not purely technical; they are socio-technical, requiring the integrated expertise of computer scientists, ethicists, legal scholars, social scientists, and humanities experts [37]. Yet, significant barriers to such collaboration persist, including divergent terminologies, differing methodological approaches, and institutional structures that often disincentivize cross-departmental research. For instance, the concept of "bias" has distinct meanings in statistics, sociology, and law, which can lead to profound misunderstandings. Fostering more effective collaboration requires a concerted effort to build a shared lexicon and common conceptual frameworks that can bridge these disciplinary divides. Joint research initiatives, co-funded by multiple directorates, and the creation of dedicated interdisciplinary centers for AI ethics are practical steps forward. Educational programs must also be redesigned to produce a new generation of researchers and practitioners who are fluent in both technical and humanistic modes of inquiry.

The trajectory of AI development will ultimately be shaped not only by experts and regulators but also by public perception and discourse. An informed public is essential for democratic accountability and for building the societal trust necessary for the widespread, beneficial adoption of AI technologies. Currently, public understanding is often hampered by sensationalized media portrayals and a lack of accessible, nuanced

information, leading to a polarized debate characterized by either utopian hype or dystopian fear. Communicating the complex trade-offs inherent in AI ethics to a general audience is a significant challenge. Future efforts should focus on creating effective strategies for public education and fostering meaningful dialogue. This includes developing clear, jargon-free explanations of how AI systems work and the ethical dilemmas they pose. The establishment of participatory governance models, such as citizen assemblies or public consultation platforms, can directly involve the public in shaping AI policy and deployment decisions, making the development process more democratic and responsive to societal values.

Looking ahead, several open questions and promising research avenues will be critical to the future of ethical AI. The ethical implications of next-generation AI paradigms, such as neuromorphic computing that mimics the brain's structure and the potentially transformative power of quantum AI, remain largely uncharted territory. There is a pressing need to move beyond one-size-fits-all ethical frameworks towards more nuanced, context-aware approaches that can adapt to the specific requirements of different domains and cultural settings. Investigating the long-term societal impacts of widespread AI adoption, including effects on labor markets, social cohesion, and human psychology, requires longitudinal studies and complex systems modeling. A critical and underexplored area is the intersection of AI ethics and environmental sustainability, examining both the significant carbon footprint of training large-scale models and the potential for AI to help solve complex environmental problems. The development of robust, standardized metrics and benchmarks for evaluating the ethical performance of AI systems is another crucial research frontier, essential for moving from principles to auditable practice.

REFERENCES

[1] M. Luengo-Oroz, "Solidarity should be a core ethical principle of AI," *Nature Machine Intelligence*, vol. 1, no. 11, p. 494, Oct. 2019, doi: 10.1038/s42256-019-0115-3.

[2] J. Morley, L. Kinsey, A. Elhalal, F. Garcia, M. Ziosi, and L. Floridi, "Operationalising AI ethics: barriers, enablers and next steps," *AI & Society*, vol. 38, no. 1, pp. 411–423, Nov. 2021, doi: 10.1007/s00146-021-01308-8.

[3] Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389–399, Sep. 2019, doi: 10.1038/s42256-019-0088-2.

[4] L. Floridi and J. Cowls, "A unified framework of five principles for AI in society," *Harvard Data Science Review*, vol. 19, no. 1, pp. 1–14, Jun. 2019, doi: 10.1162/99608f92.8cd550d1.

[5] J. Morley, L. Floridi, L. Kinsey, and A. Elhalal, "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*, vol. 26, no. 4, pp. 2141–2168, Dec. 2019, doi: 10.1007/s11948-019-00165-5.

[6] T. Hagendorff, "The Ethics of AI Ethics: An Evaluation of Guidelines," *Minds and Machines*, vol. 30, no. 1, pp. 99–120, Feb. 2020, doi: 10.1007/s11023-020-09517-8.

[7] M. Sadek and C. Mougenot, "Challenges in Value-Sensitive AI Design: Insights from AI Practitioner Interviews," *International Journal of Human-Computer Interaction*, pp. 1–18, Dec. 2024, doi: 10.1080/10447318.2024.2439021.

[8] J. Laux, S. Wachter, and B. Mittelstadt, "Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk," *Regulation & Governance*, vol. 18, no. 1, pp. 3–32, Feb. 2023, doi: 10.1111/rego.12512.

[9] M. Figueroa-Torres, "Affection as a service: Ghostbots and the changing nature of mourning," *Computer Law & Security Review*, vol. 52, p. 105943, Feb. 2024, doi: 10.1016/j.clsr.2024.105943.

[10] N. Wang, M. Christen, and M. Hunt, "Ethical Considerations Associated with 'Humanitarian Drones': A Scoping Literature Review," *Science and Engineering Ethics*, vol. 27, no. 4, Aug. 2021, doi: 10.1007/s11948-021-00327-4.

[11] L. J. Meier, A. Hein, K. Diepold, and A. Buyx, "Algorithms for Ethical Decision-Making in the Clinic: A proof of concept," *The American Journal of Bioethics*, vol. 22, no. 7, pp. 4–20, Mar. 2022, doi: 10.1080/15265161.2022.2040647.

[12] M. Al-Kfairy, D. Mustafa, N. Kshetri, M. Insiew, and O. Alfandi, "Ethical Challenges and Solutions of Generative AI: an interdisciplinary perspective," *Informatics*, vol. 11, no. 3, p. 58, Aug. 2024, doi: 10.3390/informatics11030058.

[13] Y. K. Dwivedi *et al.*, "Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy," *International Journal of Information Management*, vol. 57, p. 101994, Aug. 2019, doi: 10.1016/j.ijinfomgt.2019.08.002.

[14] S. E. Davis, C. Dorn, D. J. Park, and M. E. Matheny, "Emerging algorithmic bias: fairness drift as the next dimension of model maintenance and sustainability," *Journal of the American Medical Informatics Association*, Mar. 2025, doi: 10.1093/jamia/ocaf039.

[15] E. Ferrara, "Fairness and Bias in Artificial intelligence: A brief survey of sources, impacts, and mitigation strategies," *Sci*, vol. 6, no. 1, p. 3, Dec. 2023, doi: 10.3390/sci6010003.

[16] L. Cheng, K. R. Varshney, and H. Liu, "Socially Responsible AI algorithms: Issues, purposes, and challenges," *Journal of Artificial Intelligence Research*, vol. 71, pp. 1137–1181, Aug. 2021, doi: 10.1613/jair.1.12814.

[17] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, Oct. 2019, doi: 10.1126/science.aax2342.

[18] S. Raisch and S. Krakowski, "Artificial Intelligence and Management: The Automation–Augmentation Paradox," *Academy of Management Review*, vol. 46, no. 1, pp. 192–210, Jan. 2021, doi: 10.5465/amr.2018.0072.

[19] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, May 2019, doi: 10.1038/s42256-019-0048-x.

[20] R. O. Weber, A. J. Johs, P. Goel, and J. M. Silva, "XAI is in trouble," *AI Magazine*, vol. 45, no. 3, pp. 300–316, Jul. 2024, doi: 10.1002/aaai.12184.

[21] S. Mohseni, N. Zarei, and E. D. Ragan, "A multidisciplinary survey and framework for design and evaluation of explainable AI systems," *ACM Transactions on Interactive Intelligent Systems*, vol. 11, no. 3–4, pp. 1–45, Sep. 2021, doi: 10.1145/3387166.

[22] G. A. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, "Secure, privacy-preserving and federated machine learning in

medical imaging," *Nature Machine Intelligence*, vol. 2, no. 6, pp. 305–311, Jun. 2020, doi: 10.1038/s42256-020-0186-1.
[23] Alobaid, T. Bonny, and M. Alrahhal, "Disruptive Attacks on Artificial Neural Networks: A systematic review of attack techniques, detection methods, and protection strategies," *Intelligent Systems With Applications*, vol. 26, p. 200529, Jun. 2025, doi: 10.1016/j.iswa.2025.200529.
[24] L. Floridi *et al.*, "AI4People—An Ethical Framework for a Good AI Society: Opportunities, risks, principles, and recommendations," *Minds and Machines*, vol. 28, no. 4, pp. 689–707, Nov. 2018, doi: 10.1007/s11023-018-9482-5.
[25] L. Dung, "The argument for near-term human disempowerment through AI," *AI & Society*, vol. 40, pp. 1–14, Apr. 2024, doi: 10.1007/s00146-024-01930-2.
[26] D. C. Vladeck, "Machines without Principals: Liability Rules and Artificial Intelligence," *Washington Law Review*, vol. 89, no. 1, pp. 117–150, Mar. 2014, [Online]. Available: https://digitalcommons.law.uw.edu/wlr/vol89/iss1/6/
[27] Z. Chen, "Ethics and discrimination in artificial intelligence-enabled recruitment practices," *Humanities and Social Sciences Communications*, vol. 10, no. 1, pp. 1–12, Sep. 2023, doi: 10.1057/s41599-023-02079-x.
[28] R. B. Freeman, "Ownership when AI robots do more of the work and earn more of the income," *Journal of Participation and Employee Ownership*, vol. 1, no. 1, pp. 74–95, Jul. 2018, doi: 10.1108/jpeo-04-2018-0015.
[29] R. Radu, "Steering the governance of artificial intelligence: national strategies in perspective," *Policy and Society*, vol. 40, no. 2, pp. 178–193, Apr. 2021, doi: 10.1080/14494035.2021.1929728.
[30] B. C. Stahl, *Artificial intelligence for a better future*, 1st ed., vol. 1. Springer Nature, 2021. doi: 10.1007/978-3-030-69978-9.
[31] F. S. De Sio and G. Mecacci, "Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them," *Philosophy & Technology*, vol. 34, no. 4, pp. 1057–1084, May 2021, doi: 10.1007/s13347-021-00450-x.
[32] M. C. Gamito and C. T. Marsden, "Artificial intelligence co-regulation? The role of standards in the EU AI Act," *International Journal of Law and Information Technology*, vol. 32, no. 1, pp. 1–11, Jan. 2024, doi: 10.1093/ijlit/eaae011.
[33] M. Jirotka and A. F. T. Winfield, "Ethical Governance is essential to building Trust in Robotics and AI Systems," *Philosophical Transactions - Royal Society. Mathematical, Physical and Engineering Sciences*, vol. 376, no. 2133, pp. 1–13, Nov. 2018, doi: 10.1098/rsta.2018.0085.
[34] C. Lahusen, M. Maggetti, and M. Slavkovik, "Trust, trustworthiness and AI governance," *Scientific Reports,* vol. 14, no. 1, pp. 1–10, Sep. 2024, doi: 10.1038/s41598-024-71761-0.
[35] T. Chajdas, "Toward a fragmented global order? Technology, trade, and development in the post-2024 world," *Perspectives on Global Development and Technology*, vol. 24, no. 1–2, pp. 119–142, Apr. 2025, doi: 10.1163/15691497-12341702.
[36] P. Tambe, P. Cappelli, and V. Yakubovich, "Artificial intelligence in Human Resources Management: challenges and a path forward," *California Management Review*, vol. 61, no. 4, pp. 15–42, Aug. 2019, doi: 10.1177/0008125619867910.
[37] P. G. R. De Almeida, C. D. D. Santos, and J. S. Farias, "Artificial Intelligence Regulation: a framework for governance," *Ethics and Information Technology*, vol. 23, no. 3, pp. 505–525, Apr. 2021, doi: 10.1007/s10676-021-09593-z.

BIOGRAPHIES OF AUTHORS

Bilal Tariq is Tenured Associate Professor in the **Department of Economics, Faculty of Business Administration**, COMSATS University Islamabad, Vehari Campus. He holds a PhD in Economics with specialization in environmental and development economics. His research interests include environmental policy, digital trade, sustainable development, and economic modeling. He can be contacted at email: bilaltariq@cuivehari.edu.pk and his ORCID is https://orcid.org/0000-0003-0800-3559.

Muhammad Rehan Ashraf is Head of Department and Assistant Professor in the Department of Computer Science, Faculty of Information Science & Technology, COMSATS University Islamabad, Vehari Campus. He holds an MS in Computer Science with specialization in data science and artificial intelligence. His research areas include machine learning, AI ethics, software engineering, and computer networks. He can be contacted at email: rehanashraf1614@gmail.com, and his ORCID is https://orcid.org/0009-0000-5755-204X.

Dr. **Umar Rashid** is serving as a Lecturer in the Department of Computer Science at COMSATS University Islamabad, Vehari Campus. He has completed his Ph.D. in Computer Science, with a research focus on Medical Image Processing, Deep Learning, and Machine Learning. He is actively involved in teaching, conducting research, and supervising student projects in these domains.
Email: umarrashid@cuivehari.edu.pk
ORCID:https://orcid.org/0009-0000-8436-2348