International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

The AI/ML Ecosystem Maturity Gap: From Algorithmic Innovation to Responsible Deployment

Anandan Sonaimuthu

Abstract: The Artificial Intelligence and Machine Learning (AI/ML) ecosystem has expanded rapidly across scientific, industrial, and governmental domains, yet its technological advancement has not been matched by equivalent progress in the governance, operationalization, and lifecycle management layers that determine whether capable models become trustworthy, deployed systems. This review examines the structural architecture of the AI/ML ecosystem across six interdependent layers, data infrastructure, algorithms, computing, software frameworks, governance, and human capital, and advances the argument that the ecosystem is characterized by a persistent maturity gap between its algorithmically advanced components and its institutionally underdeveloped deployment and governance infrastructure. Drawing on empirical evidence from peer-reviewed literature published between 2015 and 2024, this review finds that fewer than 40% of organizational machine learning projects reach production deployment, reproducibility failures affect over 30% of ML-based scientific studies, and only 19% of organizations operate mature end-to-end MLOps pipelines. The review further identifies algorithmic bias, interpretability deficits, and reactive regulatory frameworks as compounding dimensions of this gap. The principal conclusion is that closing the maturity gap requires coordinated investment in governance integration, reproducibility standards, and energy-efficient deployment infrastructure across all ecosystem layers, rather than continued concentration of resources at the algorithmic frontier alone.

Keywords: *Artificial Intelligence Ecosystem, Machine Learning Deployment, AI Governance, MLOps, Maturity Gap, Explainable AI, Reproducibility, Responsible AI, AI lifecycle, Foundation Models*

1. Introduction

Artificial Intelligence (AI) and Machine Learning (ML) have undergone a transformation of historically unprecedented scale over the past decade, transitioning from specialized research disciplines into foundational technological infrastructures that underpin economic activity, scientific discovery, and public administration across the globe. The global AI market, valued at approximately USD 142.3 billion in 2023, is projected to expand at a compound annual growth rate exceeding 36% through 2030, a trajectory that reflects not merely commercial enthusiasm but the genuine penetration of AI-enabled systems into domains as varied as clinical diagnostics, autonomous transportation, financial risk modeling, agricultural yield prediction, and national security.

Cyma Systems Inc, USA

This expansion is driven by a convergence of three enabling conditions: the availability of web-scale datasets, the commoditization of high-performance computing through cloud platforms, and the maturation of deep learning architectures capable of extracting generalizable representations from raw, unstructured data. Gill et al. (2024), surveying the landscape of modern computing across its technical, institutional, and societal dimensions in *Telematics and Informatics Reports*, characterize this convergence as a structural reorganization of the computing ecosystem itself, one in which AI is no longer a layer built upon computing infrastructure but has become inseparable from it, reshaping hardware design, energy consumption patterns, and software engineering practice simultaneously.

Yet the narrative of AI's ascendance, compelling as it is when measured in benchmark performance or market capitalization, conceals a structural tension

that this review is concerned with exposing and analyzing. The AI/ML ecosystem, understood as the interconnected network of data infrastructure, algorithms, computing platforms, software frameworks, governance structures, and human expertise that collectively enable the creation and deployment of intelligent systems, has not matured uniformly across its constituent layers. Sarker (2021), in a comprehensive survey of machine learning algorithms and their real-world applications published in *SN Computer Science*, documents the breadth of ML's deployment across more than fifteen distinct application domains and identifies a recurring pattern: algorithmic performance in controlled experimental conditions consistently exceeds the performance achieved in live operational environments, often by margins of 10% to 30% on standard metrics. This performance degradation at the boundary between research and deployment is not a random artifact; it is a systematic signal of ecosystem-level immaturity in the layers responsible for translating algorithmic capability into reliable, maintainable, and governable production systems.

The structural character of this immaturity becomes clearest when the AI lifecycle is examined end-to-end. Developing a machine learning model is a well-supported activity: frameworks such as PyTorch and TensorFlow provide mature tooling, benchmark datasets are publicly available, and the academic literature offers extensive guidance on architecture selection, optimization strategies, and evaluation protocols. Deploying that model into a production environment, integrating it with live data pipelines, monitoring its behavior under distributional shift, maintaining its performance as the world changes around it, and ensuring its decisions are interpretable and auditable are activities for which the ecosystem offers dramatically less support. Lwakatare et al. (2020), examining the application of DevOps principles to AI-enabled systems in the context of the IEEE SoftCOM conference, identify six categories of challenge unique to AI deployment that have no direct equivalent in conventional software engineering: data management complexity, model decay under distribution shift, the opacity of model decision logic, the difficulty of defining "done" for a learning system, the entanglement of model and infrastructure components, and the absence of standardized testing frameworks for AI behavior. Their analysis, grounded in case studies from industrial AI projects, demonstrates that organizations attempting to apply conventional DevOps workflows to ML systems encounter these challenges systematically, irrespective of the technical sophistication of the models involved.

This review advances a central argument: that the AI/ML ecosystem is characterized by a maturity gap, a structural divergence between the sophistication of its algorithmic and computing layers, where progress has been dramatic and well-documented, and the underdevelopment of its operationalization, governance, and lifecycle management layers, where progress has been slow, uneven, and insufficiently institutionalized. The objectives of this article are threefold. First, map the six-layer architecture of the AI/ML ecosystem and characterize the maturation status of each layer against empirical evidence. Second, to diagnose the maturity gap across its principal manifestations, deployment failure rates, reproducibility deficits, governance voids, and interpretability shortfalls. Third, to identify the research directions and institutional interventions most likely to close the gap and enable the ecosystem to deliver on its widely asserted transformative potential in a manner that is reproducible, accountable, and socially responsible. The remainder of this article is organized as follows: Section 2 establishes the conceptual framework and lifecycle definition; Sections 3 through 6 analyze the ecosystem's technology, operationalization, governance, and emerging trend layers, respectively; Section 7 synthesizes the maturity gap argument; and Section 8 proposes future research directions.

2. Conceptual Framework and Methodology

2.1 Ecosystem Layer Model

The first and foremost step to address the structural dynamics of AI/ML ecosystem is to develop a coherent conceptual framework that brings together the ecosystem's technology, institution and governance aspects as an integrated analytical architecture. This paper presents a layered architecture model to understand the structural dynamics of the AI/ML ecosystem consisting of six mutually constitutive layers: data, algorithmic, infrastructure, application, governance and ethics, and human capital and institutional layers. Despite their functional separation, all layers are operationally coupled, such that a failure in one layer constrains the whole system. Governance and

operationalization, unlike previous ecosystem views, are not external to the ecosystem, but integral structural layers of an ecosystem (an important element of the maturity gap thesis).

The layered model is analytically helpful as it highlights the asymmetries in maturation speed across the ecosystem. Empirical studies have identified that the algorithmic and infrastructure layers have matured considerably faster than the governance or lifecycle management layers of the ecosystem. A systematic literature review by Paleyes et al. (2021) shows that in both academic and industrial deployments of machine learning, failure tends to happen more in the data management, model monitoring and system integration layers and less in the machine learning model layer itself, which for a long time has received the majority of researcher's focus. This asymmetry can be seen as the operational manifestation of the maturity gap that this review seeks to characterize and explain.

Recent empirical research in this area supports our diagnosis. For example, in a grounded theory study of barriers to MLOps adoption for 12 ML practitioners, Amrit and Narayanappa (2025, Journal of Innovation & Knowledge) find that only 13% of organizational ML projects are pushed to production, much lower than the 40% reported in earlier studies. The authors identify four dimensions of MLOps challenges: organizational, technical, operational and business. They find that data scientists and ML engineers spend 60-80% of their project efforts on data, not on building models, thus further supporting the observation that the maturity gap lies mostly in the non-algorithmic layers of the stack.

In Chakraborty et al. (2024), the asymmetry described above is placed in the context of what they term a systematic mapping study of MLOps pipelines, where they find that research to date has focused extensively on the model creation pipeline despite the fact that failure occurs mainly in the data manipulation and model deployment pipelines. The mapping shows that the maturity gap is more a matter of research allocation than practitioner complaint.

2.2 Review Scope and Inclusion Criteria

This article constitutes a structured narrative review of the AI/ML ecosystem, drawing on peer-reviewed publications, conference proceedings, and technical reports published between 2015 and 2024. The review is bounded thematically by the research question: to what extent does the current AI/ML ecosystem support the transition from algorithmic innovation to responsible, reproducible, and scalable real-world deployment? Sources were selected on the basis of three inclusion criteria: (i) (i) direct relevance to at least one of the six ecosystem layers defined in Section 2.1; (ii) empirical grounding through either experimental results, case study data, or systematic review methodology; and (iii) publication in indexed venues including IEEE, ACM, Elsevier, Springer, and Nature portfolio journals. Sources concerned exclusively with narrow algorithmic benchmarking without ecosystem-level implications were excluded.

The thematic scope deliberately encompasses both technological and socio-technical dimensions of the ecosystem. This dual framing is methodologically necessary because, as Amershi et al. (2019) demonstrated through a large-scale case study at Microsoft involving 515 engineers and data scientists, AI system development in practice is not a purely technical endeavor. Their study found that approximately 40% of the identified challenges in ML-enabled system development were organizational or process-related, involving issues of team coordination, data documentation, and model maintenance, concerns that fall outside the algorithmic layer entirely. This finding validates the inclusion of institutional and governance dimensions within the analytical scope of the present review and reinforces the argument that ecosystem-level analysis is necessary for understanding AI deployment outcomes.

2.3 AI Lifecycle Definition

For the purposes of this review, the AI lifecycle is defined as the end-to-end process through which an AI system is conceived, developed, validated, deployed, monitored, and retired or retrained. This definition encompasses nine sequential yet iterative phases: (1) problem formulation, (2) data acquisition and curation, (3) data preprocessing and feature engineering, (4) model selection and architecture design, (5) model training and optimization, (6) evaluation and validation, (7) deployment and integration, (8) monitoring and performance management, and (9) model update or decommissioning. This nine-phase model synthesises the lifecycle frameworks proposed across the literature and is used consistently

throughout the paper as the referential structure against which ecosystem maturity is assessed.

A critical property of this lifecycle is its non-linearity. In practice, transitions between phases are iterative and frequently reverse-directed. A model that passes evaluation may reveal data quality failures only after deployment; a production system may require architectural redesign following distributional shift in incoming data. Sculley et al. (2015) formalized this structural complexity through the concept of hidden technical debt in machine learning systems, identifying that the actual model code constitutes only a small fraction of a production ML system's total codebase. The remainder is composed of data pipelines, serving infrastructure, configuration management, monitoring modules, and process management code. Formally, if M denotes the proportion of a production system occupied by core model code, Sculley et al. (2015) characterize the system as follows:

$$System\ Complexity\ = M + \sum_{i=1}^{n} D_i$$

where D_i represents each surrounding infrastructure component (i = data pipelines, serving layers, monitoring, configuration, etc.), and n is the total number of such components. In large-scale industrial deployments, M is typically less than 5% of the total system, meaning that over 95% of production complexity resides outside the model itself. This quantitative characterization is foundational to the maturity gap argument: research communities that optimize primarily for M while neglecting D_i are producing systems that are technically sophisticated but operationally fragile.

2.4 Positioning the Maturity Gap Within the Framework

The conceptual framework and lifecycle definition established in this section collectively provide the analytical apparatus through which the maturity gap is diagnosed across subsequent sections. The gap is defined as the structural divergence between the rate of advancement in the algorithmic and infrastructure layers of the ecosystem, where progress is measured by benchmark performance, computational efficiency, and model scale, and the rate of advancement in the governance, monitoring, and operationalization layers, where progress is measured by deployment success rates, reproducibility indices, and regulatory compliance. Paleyes et al. (2021) report that fewer than 40% of organizations successfully deploy ML models into production at scale, a figure that has remained persistently low despite dramatic algorithmic progress over the same period. Amershi et al. (2019) corroborate this structural imbalance, noting that model development occupies only a portion of a typical ML engineering team's effort, with the majority of time consumed by data engineering, system integration, and post-deployment maintenance activities, precisely the D_i components that Sculley et al. (2015) identify as the true locus of production complexity. Together, these three empirical anchors establish the evidentiary foundation for the ecosystem maturity gap thesis that this review advances.

3. Technology Layers of the AI/ML Ecosystem

3.1 Data Infrastructure

The first layer of the AI/ML stack is the data. The quality, quantity, and governance properties of data determine the quality and robustness of machine learning systems. . The AI/ML stack data layer consists of data collection systems (IoT sensor networks, enterprise databases, and web-scale scraping pipelines), data lakes, data preprocessing pipelines, data labeling workflows, and data governance systems. The scale at which modern AI systems operate places extraordinary demands on this infrastructure: large language models such as GPT-3 were trained on datasets exceeding 570 gigabytes of filtered text, while vision foundation models routinely require hundreds of millions of labeled image samples. Despite this scale, the dominant sources of AI system failure are not algorithmic but data-related, with industry surveys consistently reporting that data engineers spend between 60% and 80% of their project time on data preparation, cleaning, and validation tasks rather than on modeling activities.

Renggli et al. (2021) formalize this relationship through a data quality-driven view of MLOps, arguing that model performance P is not an isolated function of architecture choice but is jointly conditioned on data quality Q and pipeline integrity I, such that:

$$P = f(Q, I, \theta)$$

where θ represents model parameters. Their analysis demonstrates that even marginal improvements in upstream data quality, measured across dimensions

of completeness, consistency, timeliness, and accuracy, produce downstream gains in model performance that are comparable in magnitude to significant architectural improvements. This quantitative framing has direct implications for ecosystem design: it repositions the data infrastructure layer from a logistical prerequisite to a primary lever of AI system quality, one that the ecosystem has not yet institutionalized with sufficient rigor.

Renggli et al. (2021) analytically derive the above relationship, which results from their data quality-driven view of MLOps, in which model performance P is not simply a function of architectural choices but the joint conditioning of data quality Q and pipeline integrity I: $P = f(Q, I, \theta)$ where θ are model parameters. Based on this relationship, the authors show that small improvements in upstream data quality in terms of completeness, consistency, timeliness and accuracy yield downstream gains in model performance similar to large architectural improvements.

Liu et al. (2025) agree in IEEE Communications Magazine, writing that "most ML-based solutions have yet to receive large-scale deployment due to insufficient maturity for production settings", and that data complexity is one of the most prominent issues. They observe that data in the networking domain is high variety (packet based, flow based, system logs, alarms, and events) and high velocity and high volume, which makes multi-modal data processing very challenging. The authors show that, at the system level, 60% of AI/ML project time is spent on maintaining data quality, which matches the literature and acts as a quantitative anchor point for the data infrastructure layer maturity gap.

3.2 Algorithms and Model Architectures

The algorithmic layer of the AI/ML ecosystem has undergone its most transformative period in the post-2017 era, driven principally by the rise of the Transformer architecture and its generalization across modalities. The major ML paradigms operating within this layer are supervised learning, unsupervised learning, reinforcement learning, and deep learning, have each advanced significantly, but the Transformer-based family of models has emerged as the dominant architectural paradigm across natural language processing, computer vision, audio processing, and multimodal tasks. Khan et al. (2021) conducted a systematic survey of vision transformers. They report over 40 model variants released in 2020-2021. Self-attention now exceeded convolutions as the state of the art (SOTA) on 11 out of 14 core computer vision datasets as of mid-2021. On ImageNet, ViT-G/14 had a top-1 accuracy of 90.45%, a number that would have been considered impossible to reach on CNN architectures as recently as in 2019.

The pace of algorithmic innovation, however, introduces its own ecosystem-level challenge. The rapid proliferation of model architectures creates substantial selection complexity for practitioners, who must navigate trade-offs between accuracy, latency, memory footprint, and interpretability without standardized evaluation protocols. This selection complexity is one dimension of the maturity gap: the algorithmic layer produces options at a rate that the governance and operationalization layers cannot assess, validate, or integrate in a commensurate timeframe.

3.3 Computing Infrastructure

The computing infrastructure layer of frontier AI includes hardware infrastructure and software platforms for training and inference. These include the use of GPUs (graphics processing units), TPUs (tensor processing units), cloud infrastructure, edge infrastructure, and distributed training frameworks. In general, the computer hardware requirements of frontier AI models are growing several times faster than Moore's Law (the number of transistors on a microchip doubles every two years). Between 2012 and 2022, the amount of compute used to train the best-performing models doubled every 3.4 months. It is estimated that training one instance of a single large-scale LLM, such as GPT-3, costs 3.14×10^{23} FLOPs, or 1,287 MWh (megawatt hours) of energy.

Dhar (2020), writing in Nature Machine Intelligence, calculated that training a single large neural architecture search model emitted approximately 626,155 pounds of CO_2 equivalent, which is the lifetime carbon emissions of an average American automobile. This bifurcates computing infrastructure from a technical layer of AI into a layer of sustainability and governance. The upstream carbon impact of AI training at scale represents one of the most consequential gaps in the AI ecosystem maturity gap: the capability of AI systems has far outstripped the institutional infrastructure for environmental accountability.

3.4 Software Frameworks and MLOps Pipelines

The software layer can include machine learning frameworks like TensorFlow, PyTorch, and JAX. It can also include experiment tracking, model registries, feature stores, and end-to-end MLOps platforms. MLOps is operationalizing the machine learning lifecycle, bringing across concepts and best-practices from the DevOps domain. Hence, MLOps has evolved as an intermediary layer between model building and deployment in production that is expected to automate, monitor and govern the same path from research spaces to production environments and overcome all reproducibility, versioning and monitoring issues present in manual deployment.

In a survey of data scientists, Mäkinen et al. (2021) found that model monitoring and automatic retraining were the highest priority unmet needs, outpacing things such as improving model accuracy. The authors also found that only 19% of survey participants indicated that they worked for organizations with mature end-to-end MLOps pipelines. These findings support the notion that there is a structural gap between the level of tooling and its uptake in organizational contexts.

Kreuzberger, Kühl, and Hirschl (2023) provide the most systematic architecture of MLOps components so far in their survey paper in the IEEE Access journal. They describe nine principles required for MLOps, including CI/CD automation, workflow orchestration, reproducibility, versioning, collaboration, continuous ML training and evaluation, tracking/logging of ML metadata, continuous monitoring, and feedback loops. Based on an analysis of 27 peer-reviewed scientific papers, 8 expert interviews, and over a hundred tool deployments, they conclude that while technical infrastructure (Kubeflow, MLflow, TensorFlow Extended) is present, MLOps is still not well adopted. They identify three main challenges: organizational (culture, skills, and cross-discipline); ML system (design for variable workload and data volume); and operational (highly reliable automation, artifacts governance & versioning, and support request resolution). This tripartite challenge framework directly operationalizes the maturity gap thesis.

In 2025, Woźniak, Milczarek, and Woźniak performed a systematic literature review on the components, tools, processes and metrics of MLOps, mapping MLOps tools to architectural components. The most frequently observed components of MLOps in the 41 papers surveyed were Model Repository (22 occurrences), Model Orchestrator (20), and Feature Store (19). The authors note that while the concept of a Feature Store is well known, only the Feast implementation is discussed in the literature and attribute this to conceptual maturity ahead of implementations, which they observe to be a common phenomenon across the maturity gap. They were surprised that no papers described metrics for evaluating the effectiveness of MLOps implementations, which they suggest is likely due to the immaturity of this aspect of the discipline. The metric vacuum is also an indicator of how underdeveloped the governance layer still is.

Adding empirical elements to similar grounded theory studies of MLOps challenges, Amrit and Narayanappa (2025) conducted 12 semi-structured interviews with ML practitioners across publishing, technology, EdTech, software, banking, consulting, and insurance. They created a typology of MLOps challenges organized under four aggregate dimensions:

- Organizational challenges: onboarding and skill shortages in ML engineers take 6-7 months, user adoption and resistance to change, long approval processes, and slow collaboration and communication between siloed teams
- Infrastructure and data management complexity, lack of standardization for MLOps tools, integration challenges with existing technical tools
- Operational difficulty: the automation of the pipelines, the trade-off between cost and prediction time, as well as over-engineering.
- Business challenges: how to communicate value to management, how to fit into the budget.

With Amrit and Narayanappa's findings in mind, we can see that the topics well-known in AI research literature like model versioning, scalability, and data drift, received less attention from practitioners than integrating tools into existing tools and infrastructure, managing data privacy, and standardization challenges. Thus, the maturity gap presents itself as a structural issue requiring intervention from the AI ecosystem.

3.5 Institutional and Industrial Actors

The AI/ML ecosystem includes universities, technology companies, government research institutes, open-source groups, and venture-backed start-ups. These institutional and industrial actors each contribute uniquely to how the AI/ML ecosystem is structured based on their respective goals, how they collaborate with each other, and what resources they make available. AI research production is highly concentrated: a 2023 study of publication patterns found fewer than ten institutions (Stanford University, Massachusetts Institute of Technology, Carnegie Mellon University, Google, and Microsoft) behind most of the most highly cited works in AI, opening up questions about structural diversity, accessibility, and geographical distribution of AI capability. Open-source tools such as Hugging Face's model hub have partially democratized these capabilities, with over 120,000 public pre-trained models uploaded in 2023. However, the computing resources to train foundation models on scale can only be accessed by a small number of well-capitalized actors. This creates an asymmetry between those who create the most powerful entities in the ecosystem and those who consume them. This form of institutional asymmetry is a socio-technical aspect of the maturity gap, requiring further consideration in governance frameworks.

4. The Maturity Gap: Where the Ecosystem Stalls

4.1 Research-to-Deployment Disconnect

The central diagnostic claim of this review is that the AI/ML ecosystem exhibits a structural maturity gap, a pronounced and persistent divergence between the sophistication of its algorithmic and infrastructure layers and the readiness of its operationalization, governance, and lifecycle management layers. This gap is not a transient artifact of a rapidly developing field; it is an embedded structural condition that manifests repeatedly across deployment contexts, institutional settings, and geographic regions. Despite benchmark performance on tasks such as image classification, natural language inference, and protein structure prediction reaching near-human or superhuman levels, real-world deployment success rates remain disproportionately low. Industry-wide estimates consistently indicate that fewer than 40% of machine learning models developed in organizational contexts are successfully transitioned into production systems, a figure that has remained stubbornly stable even as algorithmic capabilities have expanded by orders of magnitude over the same period.

The first and most detailed formal definition of this disconnect was provided by Sculley et al. (2015) when analyzing hidden technical debt in machine learning systems. In a survey of production ML systems at Google, they showed that model code itself is a small part of the ML system code, most of which is spread throughout data verification pipelines, feature engineering modules, serving infrastructure, configuration management systems, and process management code. This led to a definition of a debt ratio: sustainable systems have low ratios of the amount of boundary-crossing entanglement between the ML components and their surrounding infrastructure.

Liu et al. (2025) provide an updated analysis in light of current networking technologies, underscoring that "traditional version control tools cannot sufficiently capture the nuances of ML workflows' datasets, parameters, and configuration dependencies." Liu et al. also discuss the lack of established reproducibility workflows and the lack of cross-pollination between the different data science and network engineering priorities and domains of expertise as inefficiencies that increase time-to-value. This system-level analysis extends Sculley et al.'s debt framework by stressing the role of organizational coordination in addition to the customarily stressed debt in code.

Sallou, Durieux and Panichella (2024) identify and describe another kind of technical debt in software engineering studies that utilize Large Language Models (LLMs): the leakage of data between LLM training data and defect data used for benchmarking. They show evidence that ChatGPT has learned the defects in the widely used Defects4J defect dataset, and this raises issues for construct validity (training and evaluating on the same dataset) , and for external validity (whether results hold for unknown projects). These conclusions generalize the reproducibility crisis described by Kapoor et al. (2022) to LLM-based SE research. To reduce these threats, Sallou et al. (2023) suggest that researchers evaluate LLMs with metamorphic data (i.e., data that is transformed in a semantically-preserving manner), use multiple independent prompts, and provide execution metadata. However, the mere

existence of these threats and goal pursuits is evidence that this ecosystem has not yet operationalized responsible evaluation practices.

4.2 Lifecycle Operationalisation Failures

Operationalization failures occur at multiple phases of the AI lifecycle defined in Section 2.3, but are concentrated most heavily in phases seven through nine, deployment, monitoring, and model update. These phases receive the least attention in academic research, where publications are heavily weighted toward phases three through six (preprocessing, architecture, training, and evaluation), yet they are the phases in which the majority of practical value is either realized or lost. A persistent symptom of this imbalance is the widespread absence of systematic model monitoring in production environments: without continuous tracking of data distribution drift, prediction confidence degradation, or output fairness shifts, deployed models silently accumulate performance deficits that are only detected after consequential failures.

The challenge is complicated by the facts that most successful deep learning models are not readable or interpretable. Arrieta et al. (2020) present a taxonomy with over 400 XAI papers along two axes (types of transparency, types of explanations), forming a 2-dimensional matrix (table) of explanations. Post-hoc explanation methods such as LIME, SHAP, and gradient-based saliency maps are common in machine learning practice but rarely operationalized in production monitoring pipelines, raising questions of regulatory compliance, stakeholder trust, and error diagnosis. According to Arrieta et al. (2020), interpretability is a fundamental property in any responsible artificial intelligence application, especially in high-risk areas such as clinical decision support systems, criminal justice or financial lending.

Varga (2024), in a paper that showed the real-life impact of such operationalization failures, repeated a study by IEEE Sensors Journal that claimed to achieve 99.9% precision in human action recognition using CSI from WiFi. Upon review, Varga noted that the original authors had not partitioned their CSI images for training, validation and testing by subject. Instead, the data were randomly partitioned without ensuring that the traces of the same human subject never appeared in different training, validation, and testing partitions. In his re-implementation, Varga showed that using human subject partitioning instead of random partitioning caused the precision to drop from 99.9% to 23.4% on the WiAR dataset. Thus, dramatic performance drop in evaluations due to data leakage is not outstanding. Results using evaluation protocols that are not resistant to data leakage are systematically inflated. To address the maturity gap between research and practice, Varga proposes five strategies: (1) subject-based data partitioning, (2) transparent reporting, (3) public training curves, (4) reviewer scrutiny of data partitioning decisions, and (5) publisher guidance.

4.3 Reproducibility and Benchmarking Deficits

While all three elements contribute to the maturity gap, the reproducibility crisis is the most damaging in terms of its structural impact. Kapoor and Narayanan (2022) found in a systematic survey of 17 scientific domains where ML approaches are adopted that data leakage is the leading cause of incorrect performance estimates in the literature that dictate the course of research. They found leaks in over 30% of studies in biomedicine, economics, and climate science, and several leaks in all three domains resulted in productivity increases of 3% to 100% above leak-free baselines. This means that benchmarks often cited as evidence of AI maturity are at least partially, and often completely, methodologically nonrobust. The ecosystem is characterized by a perceived capability frontier that is considerably out of step with deployable and deployed systems. This reproducibility crisis is not incidental but systemic, caused by a lack of standardization in data partitioning, weak experimental peer review practices, and academic publishing practices that prioritize novelty over reproducibility and robustness.

Table 1: Key Numerical Metrics of the AI/ML Ecosystem Maturity Gap

Metric	Value
Organisational ML projects reaching production	< 40%
ML studies affected by data leakage	> 30%

Organisations with mature MLOps pipelines	19%
ML engineering effort spent outside model development	~60%
Model code as proportion of production system	< 5%
CO_2 equivalent from one NAS model training run	626,155 lbs
Performance degradation: research to deployment	10–30%

Note. Metrics are drawn directly from cited empirical studies and are used as quantitative anchors for the maturity gap thesis throughout this review.

5. Governance, Ethics, and Regulatory Dimensions

5.1 Algorithmic Bias and Fairness

The governance layer of the AI/ML ecosystem has emerged as one of the most consequential and least institutionally developed strata of the entire system. As AI applications penetrate high-stakes domains including criminal justice, healthcare, financial lending, and employment screening, the societal consequences of governance failures have become measurable and documentable. Algorithmic bias, the systematic production of discriminatory outputs by ML models as a result of skewed training data, flawed problem formulation, or proxy variable entanglement, represents the most extensively studied governance failure mode. Empirical evidence of its real-world impact is substantial: the COMPAS recidivism prediction system, deployed across multiple US jurisdictions, was found to misclassify Black defendants as high-risk at approximately twice the rate applied to white defendants, despite achieving comparable overall accuracy. Similarly, facial recognition systems evaluated by the National Institute of Standards and Technology in 2019 exhibited false positive rates up to 100 times higher for darker-skinned females than for lighter-skinned males across 189 tested algorithms.

A survey by Caton and Haas (2020) collected the 60+ mathematical definitions of algorithmic fairness proposed in the academic literature, categorizing them into the three core categories of individual fairness, group fairness, and causal fairness. The authors identify a fundamental conflict behind fairness definitions: if the base rates between groups differ, many of the definitions are provably incompatible.

The most thorough recent overview of the "principle proliferation" problem can be found in Gunasekara et al. (2025) 's systematic review of responsible AI principles and practice, published in MDPI Applied System Innovation. Out of 22,711 publications initially searched, 553 peer-reviewed publications were included in the final corpus. The review identifies seven principles of responsible AI: transparency and explainability; fairness and algorithmic bias; privacy and data protection; robustness and reliability; accountability; human agency and oversight; and socially helpful practice. The review finds principles are ubiquitous, but little evidence was found of efforts to translate principles into practice. For example, they find that few studies focus on ex ante accountability approaches to prevent harm before system deployment, while most accountability frameworks are ex post retrospective attribution mechanisms that focus on assessing harm after system deployment. This can also be interpreted as support for the governance layer immaturity diagnosis, as Gunasekara et al. also comment on the difference between responsible AI, ethical AI, and trustworthy AI, and argue that the term responsible AI is the more accurate designation as it reduces the ambiguity of ethical and trustworthy AI. A BERTopic analysis of the 553 articles identified 15 main topics, the three most common being "AI in Healthcare and Digital Medicine" (22%), "Responsible AI Principles and Stakeholder Governance" (11.2%), and "ChatGPT and Academic Integrity in Education" (9.7%).

5.2 Ethical Frameworks and Responsible AI Principles

In addition to the operational dimensions of bias and fairness, the design and governance of the AI/ML ecosystem could be guided by ethical principles. From a multi-disciplinary consensus of ethicists, lawyers, computer scientists, and policymakers, the AI4People framework by Floridi et al. (2018) proposes a foundation for the design, deployment, and regulation of AI systems around the five principles of good AI society: beneficence, non-maleficence, autonomy, justice, and explicability.

Of special relevance was its success in assembling for the first time bioethics, human rights law, and AI safety research in a coherent governance architecture for organisations and societies, comprising 47 concrete policy recommendations across the four domains. This AI4People consensus report was critical in showing that the responsible governance of AI cannot be reduced to a purely technical exercise, but requires concerted institutional action at all levels of the ecosystem. This theoretical framework has subsequently followed into practice in two nascent regulatory measures: the European Commission's Ethics Guidelines for Trustworthy AI (2019) and the OECD Principles on Artificial Intelligence.

5.3 Algorithmic Accountability and the Regulatory Landscape

Institutional instruments for holding AI systems accountable have also failed to keep up. A systematic literature review of algorithmic accountability, conducted by Wieringa (2020) and presented at the ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT 2020), found that the majority of accountability frameworks were retrospective attribution mechanisms, i.e., attempts to attribute blame ex post to harmful outcomes. Compared to these studies, there were fewer studies focused on ex ante accountability approaches to ensure systems do not cause harm.

Ferjani, Alzahrani and Bouzir's (2025) systematic review of the AI governance literature in Elsevier's Procedia Computer Science proposes a new "Governance Galaxy" framework of AI co-governing humanity by 2035. They analyzed 75 peer-reviewed articles between 2010 and 2024, identifying four clusters of ideas: ethics, regulation, technology, and global coordination. They conclude that there remains a clear gap between theoretical principles and their realization in practice, particularly with respect to "under-represented stakeholders", such as indigenous peoples, SMEs, and non-Western perspectives. A quality assessment of the studies indicated that while 65% were high quality studies (MMAT score 10-12), few studies sought to identify solutions for implementation. These include Utopian, Dystopian and Fragmented governance scenarios, as well as examples in Denmark, Canada, Saudi Arabia and EU-Asia policy dialogue. They also provide a governance toolbox, including such instruments as regulatory sandboxes or decentralized governance processes. They theorize that the gap between articulating principles and their implementation, such as the 2027 implementation date of the EU AI Act, is not merely a temporary issue, but rather a structural feature of the governance layer's immaturity.

6. Emerging Trends Reshaping the Ecosystem

6.1 Generative AI and Foundation Models

The biggest structural change to the AI/ML stack between 2020 and 2024 has been the emergence of LLMs, more broadly known as foundation models, which are generative AI models trained on web-scale data with self-supervised objectives and adapted to individual downstream tasks (i.e., applications) with prompting and fine-tuning. Foundation models have transformed the algorithmic layer described in Section 3.2, replacing task-specific model training with a new model of general-purpose pre-training on a large dataset, followed by lightweight task adaptation. These models are massive. OpenAI's GPT-3, which was released in 2020, has 175 billion parameters, and was trained on approximately 570 GB of filtered text data. Training was done with an estimated more than 3.14×10^{23} floating point operations. Models have only grown larger since GPT-3's release. Google's PaLM, which had 540 billion parameters, and Meta's LLaMA 2, which showed results at 70 billion parameters, suggested that high-quality performance could be maintained with fewer parameters, which has implications for sustainability and accessibility.

Zhao et al. (2023) conducted the most comprehensive survey of large language models available in the literature, reviewing over 250 LLMs developed between 2018 and 2023 and systematically analyzing their pre-training data, architectural configurations, fine-tuning strategies, and evaluation methodologies. Their analysis identifies four emergent capabilities that arise in LLMs above a threshold of approximately 10 billion parameters and that are not present in smaller models: in-context learning, instruction following, step-by-step reasoning, and tool use. The emergence of these capabilities at scale represents a qualitative shift in the ecosystem's algorithmic layer, one that simultaneously expands deployable capability and deepens the governance challenges identified in Section 5, since the outputs of LLMs are

considerably harder to audit, explain, or constrain than those of narrower task-specific models. The generative AI trend, therefore, does not resolve the maturity gap; it widens it by introducing systems whose societal footprint grows faster than the institutional capacity to govern them.

6.2 Automated Machine Learning and Democratisation

The second major trend reshaping the ecosystem is the maturation of Automated Machine Learning (AutoML), the systematic automation of model selection, hyperparameter optimization, feature engineering, and neural architecture search, and its role in lowering the technical barriers to AI adoption. AutoML addresses a structural bottleneck in the human capital layer of the ecosystem: the global shortage of qualified ML engineers and data scientists, estimated by the World Economic Forum at over one million unfilled positions by 2022, creates a deployment ceiling that cannot be resolved through workforce expansion alone. By automating the most technically demanding phases of the ML lifecycle, AutoML tools extend productive AI capability to domain experts in medicine, agriculture, law, and education who possess deep subject knowledge but limited algorithmic training.

He et al. (2021) published a comprehensive survey of the AutoML state of the art in *Knowledge-Based Systems*, cataloging methods across four primary automation domains: hyperparameter optimization, neural architecture search (NAS), meta-learning, and full-pipeline automation. Their analysis documents that NAS methods, which search over discrete spaces of architectural configurations to identify optimal designs, had by 2020 produced models outperforming human-designed architectures on ImageNet while reducing the expert time required for architecture development from months to hours. Formally, the NAS problem can be expressed as:

$$\alpha^* = arg\ max_{\alpha \varepsilon A}\ val - acc(w^*(\alpha)\ D_{val})$$

where α denotes an architecture sampled from search space A, $w(\alpha)$ are the optimized weights for that architecture trained on D_{train}, and val-acc measures validation accuracy on D_{val}. He et al. (2021) report that efficient NAS approaches such as DARTS reduced GPU search time from over 2,000 GPU-days under early reinforcement learning-based methods to fewer than four GPU-days, a reduction of greater than 99.8%, making architecture search computationally accessible to organizations without hyperscale computing resources. Despite this progress, AutoML adoption remains uneven across the ecosystem: the same governance and operationalization deficits that characterize manually developed models apply with equal force to AutoML-generated pipelines, and the opacity of automatically designed architectures introduces additional interpretability challenges that the governance layer is not yet equipped to address systematically.

6.3 Edge AI, Human-Centred Design, and Ecosystem Integration

Two further trends complete the picture of an ecosystem in active structural transition. Edge AI, the deployment of inference-capable ML models on resource-constrained devices including smartphones, industrial sensors, and autonomous vehicles, extends the application layer of the ecosystem into environments where cloud connectivity is unreliable, latency constraints are stringent, and data privacy considerations preclude centralized processing. Estimates from IDC projected that by 2025, approximately 75% of enterprise-generated data would be created and processed outside traditional centralized data centers, a trajectory that makes edge inference not a peripheral capability but a mainstream deployment requirement. The technical challenge of edge AI is model compression: achieving acceptable task performance within the memory, compute, and energy budgets of edge devices. hardware through techniques including quantization, pruning, and knowledge distillation. Concurrently, human-centered AI design, the integration of human factors, participatory design methods, and user mental model research into the AI development lifecycle have gained significant traction as a counterweight to purely performance-driven development paradigms. Together, edge AI and human-centered design represent the ecosystem's most active sites of expansion into the D_i infrastructure components identified by Sculley et al. (2015), extending not the model itself but the surrounding system that makes the model usable, trustworthy, and contextually appropriate.

Table 2: Emerging Trends and Their Governance Implications for the Ecosystem

Trend	Capability Advance	Governance Challenge Introduced
Large Language Models	Emergent reasoning above 10B parameters (Zhao et al., 2023)	Outputs harder to audit, explain, or constrain
AutoML and NAS	GPU search time reduced by > 99.8% via DARTS (He et al., 2021)	Automated architectures deepen interpretability deficit
Edge AI	75% of enterprise data generated outside data centres by 2025	Data privacy and model compression standards absent
Generative AI	GPT-3: 175B parameters; PaLM: 540B parameters	Societal footprint grows faster than regulatory capacity
AI Democratisation	Hugging Face: > 120,000 public pre-trained models by 2023	Governance tools not accessible to non-expert deployers

Note. Each trend expands the ecosystem's capability frontier while simultaneously introducing governance demands that the current institutional layer is not equipped to meet, compounding rather than closing the maturity gap.

7. Discussion

7.1 Synthesising the Maturity Gap Thesis

The analysis from Sections 3 to 6 thus converge on a unifying structural diagnosis of the AI/ML ecosystem: an enduring and deep asymmetry between its technologically advanced upstream layers and its institutionally undersophisticated governance, operationalization, and lifecycle management layers. This maturity gap is not simply a temporary state due to the relative youth of the AI/ML field. This is a property that persists across deployment contexts, institutional contexts, and generations of models, because research incentives, commercial competition, and regulatory design are all focused more on the capabilities of models than their deployment quality.

The technical debt framework proposed by Sculley et al. (2015) is one of the most stable perspectives to study technical debt. Amrit & Narayanappa (2025) validate this empirically, citing the onboarding time of ML engineers (6-7 months), communication breakdowns between teams and the absence of institutionalized testing regimes for non-deterministic ML systems as enduring organizational-level technical debts. More recently, Kreuzberger, Kühl and Hirschl (2023) show, that MLOps maturity does not only mean adoption of technical components but also changing culture towards product in machine learning (product-oriented discipline) and that most organizations are still far away from reaching this stage.

As Woźniak et al. (2025) summarize this part, as there is no standard metric to assess MLOps implementation effectiveness, organizations cannot measure their MLOps maturity level and cannot compare themselves to their peers. The lack of such metrics, however, is in itself a problem of underdevelopment at the level of the governance layer.

7.2 Implications for Research and Practice

The governance dimension of the maturity gap carries implications that extend beyond engineering practice into public policy and institutional design. Arrieta et al. (2020) establish that interpretability is a prerequisite for responsible deployment in high-stakes domains, yet their survey of over 400 XAI papers finds that post-hoc explanation methods remain overwhelmingly absent from production monitoring pipelines. This absence is not technical; the tools exist, but institutionally, organizations lack the regulatory pressure, internal expertise, and process infrastructure to operationalize explainability systematically. Floridi et al. (2018) anticipated this institutionalization gap through the AI4People framework, arguing that the five principles of beneficence, non-maleficence, autonomy, justice, and explicability require coordinated implementation across individual, organizational, and governmental levels

simultaneously. The evidence reviewed in Section 5 suggests that this coordinated implementation has not materialized at the required scale: regulatory frameworks such as the EU AI Act impose compliance obligations on a timeline that extends to 2027 for high-risk systems, meaning that the governance layer continues to operate reactively rather than prospectively.

7.3 Towards an Integrated Ecosystem Maturity Model

Resolving the maturity gap requires a shift in how the AI/ML community conceptualizes and measures progress. The field currently lacks a standardized ecosystem maturity model, analogous to the Capability Maturity Model Integration (CMMI) used in software engineering, that assesses AI systems not only on algorithmic performance but also on data governance quality, lifecycle management completeness, reproducibility standards, and ethical compliance. Mäkinen et al. (2021) find that only 19% of organizations operate with mature end-to-end MLOps pipelines, suggesting that the majority of AI-deploying organizations are operating at maturity levels one or two on any reasonable five-level scale. Caton and Haas (2020) demonstrate that fairness cannot be reduced to a single metric, implying that a maturity model for governance must accommodate plural, context-sensitive fairness criteria rather than universal compliance thresholds. An integrated ecosystem maturity model that spans all six layers defined in Section 2.1, from data infrastructure through institutional governance, would provide researchers, practitioners, and regulators with a shared evaluative framework capable of diagnosing where the gap is widest and directing investment accordingly. Developing, validating, and institutionalizing such a model represents the most consequential single contribution the research community could make to closing the maturity gap.

8. Future Research Directions

8.1 Trustworthy and Explainable AI Systems

The most urgent research priority emerging from the maturity gap analysis is the development of trustworthy AI systems, systems that are not only accurate but verifiably reliable, interpretable, and aligned with human values across their operational lifetime. Current explainability research, as documented by Arrieta et al. (2020) across more than 400 reviewed papers, has produced a rich taxonomy of post-hoc explanation methods but has made comparatively limited progress on ante-hoc interpretability, the design of model architectures that are intrinsically transparent rather than explained retrospectively. Future research must therefore prioritize the development of high-performance architectures that embed interpretability as a design constraint rather than an afterthought and must develop standardized evaluation protocols for explanation quality that go beyond fidelity metrics to assess whether explanations are actionable, stable, and comprehensible to non-expert stakeholders. The clinical AI domain, where model decisions directly influence patient outcomes, provides the most stringent test environment for this research agenda: studies have shown that AI diagnostic systems achieving area-under-curve scores exceeding 0.95 on held-out test sets have nonetheless failed regulatory approval due to the absence of interpretable decision pathways, illustrating the practical cost of the interpretability deficit.

8.2 Energy-Efficient and Sustainable Machine Learning

The computational trajectory documented by Dhar (2020), in which training a single large neural architecture search model produces approximately 626,155 pounds of CO_2 equivalent, makes energy efficiency a research imperative rather than a secondary engineering concern. The carbon cost of frontier model training is growing at a rate that is incompatible with national and international climate commitments, and the democratization trend documented by He et al. (2021) through AutoML will amplify aggregate energy consumption by enabling a larger population of actors to train large models. Future research must therefore pursue advances on three interconnected fronts: first, the development of training algorithms that achieve equivalent generalization performance with substantially fewer FLOPs, including sparse training, curriculum learning, and early exit architectures; second, the design of hardware-software co-optimized inference pipelines for edge deployment that reduce energy per inference by orders of magnitude relative to cloud-based serving; and third, the establishment of standardized energy reporting requirements for published ML research, analogous to the financial reporting standards that make corporate energy consumption auditable.

Without the third advance, the first two cannot be systematically incentivized or verified.

9. Conclusion

This review has demonstrated that the AI/ML ecosystem is characterized by a structural maturity gap, a persistent and consequential divergence between the sophistication of its algorithmic and computing layers and the underdevelopment of its operationalization, governance, and lifecycle management layers. Despite benchmark achievements of considerable scale, including vision transformer models exceeding 90% top-1 accuracy on ImageNet and large language models demonstrating emergent reasoning capabilities above ten billion parameters, fewer than 40% of organizational ML projects reach production deployment, and reproducibility failures affect over 30% of ML-based scientific studies across seventeen domains. The ecosystem's model code constitutes fewer than 5% of a production system's total complexity, yet research investment remains disproportionately concentrated on that fraction, while the surrounding infrastructure of data pipelines, monitoring systems, and governance mechanisms, the components that determine whether capable models become trustworthy deployed systems, remains structurally underinvested. Progress measured solely at the algorithmic frontier is therefore a misleading indicator of ecosystem maturity, and the field requires a unified maturity assessment framework spanning all six ecosystem layers identified in this review to make the gap visible, measurable, and addressable.

Closing the maturity gap demands coordinated action across three interdependent fronts. Institutionally, governance frameworks must be embedded into the AI lifecycle from problem formulation onward rather than appended reactively after deployment failures have occurred, with regulatory instruments such as the EU AI Act calibrated through empirical regulatory science rather than the precautionary principle alone. Technically, future research must prioritize energy-efficient training, ante-hoc interpretability architectures, and standardized MLOps pipelines that make reproducibility and monitoring structural properties of AI systems rather than optional engineering practices. Epistemically, the reproducibility crisis must be addressed through mandatory data partitioning standards and pre-registration norms that align the incentive structures of academic publication with the rigor that responsible deployment demands. The AI/ML ecosystem possesses the raw capability to contribute meaningfully to healthcare. climate science, education, and public administration, but realizing that potential depends not on further algorithmic innovation alone but on the collective commitment of researchers, practitioners, and regulators to invest in the layers that transform capability into trustworthy, governable, and sustainable deployment.

References

[1] Gill, S. S., Wu, H., Patros, P., Ottaviani, C., Arora, P., Pujol, V. C., Haunschild, D., Parlikad, A. K., Cetinkaya, O., Lutfiyya, H., Stankovski, V., Li, R., Ding, Y., Qadir, J., Abraham, A., Ghosh, S. K., Song, H. H., Sakellariou, R., Rana, O., & Rodrigues, J. J. P. C. (2024). Modern computing: Vision and challenges. Telematics and Informatics Reports, 13, 100116. https://www.sciencedirect.com/science/article/pii/S2772503024000021

[2] Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications, and Research Directions. SN Computer Science, 2(3), 1–21. Springer. https://link.springer.com/article/10.1007/s42979-021-00592-x

[3] Lwakatare, L. E., Crnkovic, I., & Bosch, J. (2020, September 1). DevOps for AI – Challenges in Development of AI-enabled Applications. IEEE Xplore. https://doi.org/10.23919/SoftCOM50211.2020.9238323

[4] Paleyes, A., Urma, R.-G., & Lawrence, N. D. (2021). Challenges in Deploying Machine Learning: a Survey of Case Studies. ArXiv:2011.09926 [Cs]. https://arxiv.org/abs/2011.09926

[5] Amershi, S., Begel, A., Bird, C., Deline, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B., & Zimmermann, T. (n.d.). Software Engineering for Machine Learning: A Case Study. https://www.microsoft.com/en-us/research/wp-content/uploads/2019/03/amershi-icse-

2019_Software_Engineering_for_Machine_Learning.pdf

[6] Renggli, C., Rimanic, L., Gürel, N. M., Karlaš, B., Wu, W., & Zhang, C. (2021). A Data Quality-Driven View of MLOps. ArXiv:2102.07750 [Cs]. https://arxiv.org/abs/2102.07750

[7] Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2021). Transformers in Vision: A Survey. ArXiv:2101.01169 [Cs]. https://arxiv.org/abs/2101.01169

[8] Dhar, P. (2020). The Carbon Impact of Artificial Intelligence. Nature Machine Intelligence, 2(8), 423–425. https://doi.org/10.1038/s42256-020-0219-9

[9] Mäkinen, S., Skogström, H., Laaksonen, E., & Mikkonen, T. (2021). Who Needs MLOps: What Data Scientists Seek to Accomplish and How Can MLOps Help? ArXiv:2103.08942 [Cs]. https://arxiv.org/abs/2103.08942

[10] Kapoor, S., & Narayanan, A. (2022). Leakage and the Reproducibility Crisis in ML-Based Science. ArXiv:2207.07048 [Cs, Stat]. https://arxiv.org/abs/2207.07048

[11] Arrieta, A. B., Díaz-Rodríguez, N., Ser, D., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2019). Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities, and Challenges toward Responsible AI. ArXiv.org. https://arxiv.org/abs/1910.10045

[12] Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden Technical Debt in Machine Learning Systems. https://proceedings.neurips.cc/paper_files/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf

[13] Caton, S., & Haas, C. (2020). Fairness in Machine Learning: A Survey. ArXiv:2010.04053 [Cs, Stat]. https://arxiv.org/abs/2010.04053

[14] Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People, An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds and Machines, 28(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

[15] Wieringa, M. (2020). What to account for when accounting for algorithms. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 1–18. https://doi.org/10.1145/3351095.3372833

[16] Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., & Liu, P. (2026). A Survey of Large Language Models. ArXiv:2303.18223 [Cs]. https://arxiv.org/abs/2303.18223

[17] He, X., Zhao, K., & Chu, X. (2021). AutoML: A survey of the state-of-the-art. Knowledge-Based Systems, 212, 106622. https://doi.org/10.1016/j.knosys.2020.106622

[18] Liu, Q., Zhang, T., Masoud Hemmatpour, Qiu, H., Zhang, D., Chen, C. S., Mellia, M., & Armen Aghasaryan. (2024). Operationalizing AI/ML in Future Networks: A Bird's Eye View from the System Perspective. IEEE Communications Magazine, 1–7. https://doi.org/10.1109/mcom.001.2400033

[19] Chakraborty, A., Das, S., & Gary, K. (2025). Machine Learning Operations: A Mapping Study. Communications in Computer and Information Science, 3–21. https://doi.org/10.1007/978-3-031-86644-9_1

[20] Amrit, C., & Narayanappa, A. K. (2024). An analysis of the challenges in the adoption of MLOps. Journal of Innovation & Knowledge, 10(1), 100637. https://doi.org/10.1016/j.jik.2024.100637

[21] Varga, D. (2024). Critical Analysis of Data Leakage in WiFi CSI-Based Human Action Recognition Using CNNs. Sensors, 24(10), 3159. https://doi.org/10.3390/s24103159

[22] Sallou, J., Durieux, T., & Panichella, A. (n.d.). Breaking the Silence: the Threats of Using LLMs in Software Engineering. https://doi.org/10.1145/3639476.3639764

[23] Woźniak, A. P., Milczarek, M., & Woźniak, J. (2025). MLOps Components, Tools, Process and Metrics - A Systematic Literature Review. IEEE Access, 1–1. https://doi.org/10.1109/access.2025.3534990

[24] Kreuzberger, D., Kühl, N., & Hirschl, S. (2023). Machine Learning Operations (MLOps): Overview, Definition, and Architecture. IEEE

Access, 11, 1–1. https://doi.org/10.1109/access.2023.3262138

[25] Daochen Zha, Bhat, Z. P., Lai, K.-H., Yang, F., Jiang, Z., Zhong, S., & Hu, X. (2025). Data-centric Artificial Intelligence: A Survey. ACM Computing Surveys. https://doi.org/10.1145/3711118

[26] Lakshitha Gunasekara, El-Haber, N., Nagpal, S., Harsha Moraliyage, Zafar Issadeen, Milos Manic, & Silva, D. D. (2025). A Systematic Review of Responsible Artificial Intelligence Principles and Practice. Applied System Innovation, 8(4), 97–97. https://doi.org/10.3390/asi8040097

[27] Ferjani, I., Alzahrani, F. A., & Bouzir, N. M. (2025). Navigating Responsible AI: A Systematic Review of Governance Mechanisms and Future Co-Governance Scenarios. Procedia Computer Science, 270, 4726–4735. https://doi.org/10.1016/j.procs.2025.09.598

[28] Manal Alghieth. (2025). Sustain AI: A Multi-Modal Deep Learning Framework for Carbon Footprint Reduction in Industrial Manufacturing. Sustainability, 17(9), 4134–4134. https://doi.org/10.3390/su17094134