Accepted in *AI and Ethics* (2026) here

Building the Ethical AI Framework of the Future: From Philosophy to Practice

Jasper Kyle Catapang[1,2*]

[1*]Graduate School of Global Studies
Tokyo University of Foreign Studies, Tokyo, Japan .
[2*]AI Product Development Division
Money Forward, Inc., Tokyo, Japan .

Corresponding author(s). E-mail(s): catapang.jasper.kyle.y0@tufs.ac.jp;

Abstract

Artificial intelligence pipelines—spanning data collection, model training, deployment, and post-deployment monitoring—concentrate ethical risks that intensify with multimodal and agentic systems. Existing governance instruments, including the EU AI Act, the IEEE 7000 series, and the NIST AI Risk Management Framework, provide high-level guidance but often lack enforceable, end-to-end operational controls. This paper presents an ethics-by-design control architecture that embeds consequentialist, deontological, and virtue-ethical reasoning into stage-specific enforcement mechanisms across the AI lifecycle. The framework implements a triple-gate structure at each lifecycle stage: Metric gates (quantitative performance and safety thresholds), Governance gates (legal, rights, and procedural compliance), and Eco gates (carbon and water budgets and sustainability constraints). It specifies measurable trigger conditions, escalation paths, audit artefacts, and mappings to EU AI Act obligations and NIST RMF functions, enabling integration with existing MLOps and CI/CD pipelines. Illustrative examples from large language model pipelines demonstrate how gate-based controls can surface and constrain technical, social, and environmental risks prior to release and during runtime. The framework is accompanied by a preregistered evaluation protocol that defines ex ante success criteria and assessment procedures, enabling falsifiable evaluation of gate effectiveness. By translating normative commitments into enforceable and testable controls, the framework provides a practical basis for operational AI governance across organizational contexts, jurisdictions, and deployment scales.

Keywords: Ethical AI, AI pipelines, Responsible AI, Philosophy and AI, Lifecycle governance, Large Language Models, Sustainable AI

1 Introduction

The rise of Artificial Intelligence (AI) has reshaped Natural Language Processing (NLP) and Information Retrieval (IR), powering applications from search and summarization to conversational agents and retrieval-augmented generation (RAG) [1]. Large Language Models (LLMs) currently exemplify this transformation, yet the field is evolving toward multimodal foundation models, agentic AI systems, and hybrid neuro-symbolic architectures [2, 3]. Regardless of the paradigm, AI pipelines follow a similar lifecycle:

1. **Data collection and curation**, which raises issues of privacy, intellectual property, and representational fairness [4];
2. **Model training and alignment**, where biases, harmful knowledge, and unbalanced datasets can propagate societal risks [5];
3. **Deployment and inference**, where hallucination, misuse, and opaque decision-making can harm users at scale [6];
4. **Post-deployment monitoring**, where drift detection, auditing, and accountability mechanisms remain underdeveloped [7].

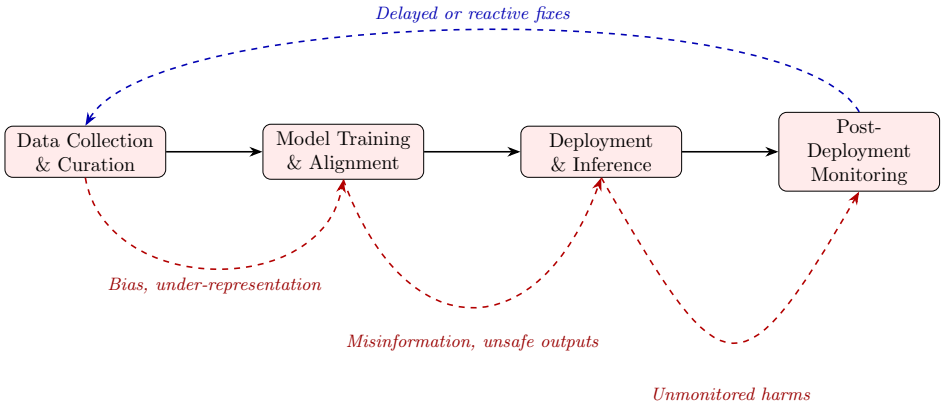


Fig. 1 How early-stage ethical failures propagate forward; monitoring feedback often arrives too late.

Figure 1 illustrates how ethical vulnerabilities can cascade across the AI development lifecycle. Issues originating during data collection and curation, such as biased sampling or under-representation, tend to persist and amplify during model training and alignment, manifesting as misinformation or unsafe outputs during deployment and inference. Weaknesses in post-deployment monitoring exacerbate the problem, as feedback loops are often delayed or reactive rather than preventative. This sequential propagation means that late-stage interventions, while necessary, are inherently less effective than proactive measures embedded at earlier stages of the pipeline.

Traditional AI ethics guidelines emphasize fairness, accountability, and transparency, yet they rarely translate into concrete engineering practices for model

pipelines [8]. Current NLP evaluation metrics (BLEU, ROUGE, Recall@k) overlook ethical dimensions, and post-hoc mitigation often lags behind real-world harms.

This paper makes three contributions. First, it proposes a control architecture that systematizes existing ethics and governance frameworks into an end-to-end operational pattern usable today and adaptable to emerging AI paradigms. Second, it formalizes the enforcement of consequentialist, deontological, and virtue-ethical commitments via stage-specific controls implemented as a triple gate (metric, governance, eco) with measurable trigger conditions, escalation paths, and auditable artefacts, and provides a crosswalk to the EU AI Act, IEEE 7000 series, and the NIST AI RMF to support conformity assessment and risk management. Third, it specifies a preregistered evaluation protocol and illustrates the framework's application to large-language-model pipelines, showing how the control structure extends to multimodal, agentic, and neuro-symbolic systems.

This work contributes a control layer that existing frameworks lack: a triple-gate at each pipeline stage with explicit, testable triggers that can be automated in CI/CD and enforced at runtime. Unlike prior lifecycle ethics approaches that emphasize principles, documentation, or post-hoc reporting, the contribution here is operational: (i) measurable thresholds that decide promotion/rollback within MLOps, (ii) governance artefacts aligned to conformity assessment, and (iii) an Eco gate that makes carbon and water budgets first-class criteria alongside accuracy and fairness. This turns ethics from an after-action audit into a gatekeeping mechanism embedded in build, deploy, and serving paths.

To position this contribution relative to existing research, this paper does not propose new ethical principles or normative theories—the consequentialist, deontological, and virtue-ethical foundations are well-established in moral philosophy—nor does it introduce new fairness metrics or evaluation methods, as the metrics referenced (demographic parity, equalized odds, KL divergence, etc.) are drawn from existing literature. Rather, the contribution is a missing abstraction layer between high-level ethical principles and operational MLOps pipelines: a formalized control architecture that enables existing principles and metrics to function as enforceable lifecycle constraints. This is analogous to how safety-critical engineering disciplines—from aviation systems to medical devices to DevSecOps practices—have developed control topologies (circuit breakers, fail-safe mechanisms, continuous integration gates) that translate design principles into operational safeguards. The framework does not claim immediate empirical superiority over existing approaches; empirical validation is intentionally separated from the design and protocol specification (see Section 6.4) to enable falsifiable evaluation without post-hoc threshold tuning.

2 Background and AI Pipeline Overview

Understanding the ethical dimensions of AI requires first unpacking how modern AI systems are built, deployed, and maintained. The notion of an *AI pipeline* captures the multi-stage process through which raw data is transformed into operational models that interact with users and society. Each stage presents distinct ethical stakes, and lapses early in the pipeline often propagate downstream in ways that are difficult to reverse.

2.1 From Classical AI to Modern Pipelines

Early AI workflows were relatively linear and interpretable. Symbolic reasoning systems and small-scale statistical models were trained on curated datasets, which constrained both their capabilities and their ethical surface area [8]. Errors were easier to trace, and risks were mostly limited to system reliability or local data misuse.

Modern AI, by contrast, operates in sprawling, data-intensive pipelines. Contemporary systems exhibit three defining trends:

- **Massive and heterogeneous datasets,** often scraped from the open web or user-generated platforms, which blur the boundaries between public and private information [2, 4];
- **Deep and opaque architectures,** such as foundation models and LLMs, capable of emergent behaviors that even their developers may not fully anticipate [3, 9];
- **Persistent and iterative lifecycles,** in which models are continuously fine-tuned, redeployed, and monitored rather than released as static products [7].

This evolution amplifies the ethical stakes: failures in early stages can cascade into systemic harm, and reactive interventions are often inadequate. Viewing AI development through a pipeline lens is thus essential for designing proactive, ethics-by-design interventions.

2.2 Data Collection and Curation

Data is the substrate of modern AI, and ethical risk begins at the point of collection. Large Language Models (LLMs) and multimodal systems ingest vast quantities of text, images, code, and audio, often without explicit user consent [2]. Three interrelated concerns dominate this stage.

First, **privacy and consent** are easily violated by indiscriminate web scraping or by incorporating user-generated data without transparent notice. Even nominally anonymized datasets may retain identifiable traces, especially for high-dimensional media like geolocated images or conversational logs. Second, **intellectual property and creative labor** are frequently implicated: generative models may train on copyrighted material or artists' work without attribution or compensation, raising questions of both legality and fairness. Third, **representation and bias** emerge as structural risks. Web-derived corpora tend to overrepresent English and Western cultural content, underrepresenting minority languages and perspectives. The result is a pipeline that can encode and reproduce global inequities [4].

Additionally, scholars highlight the phenomenon of *data colonialism*, in which cultural and linguistic resources from the Global South are extracted for AI development without local governance or reciprocal benefit [10]. Because biases and omissions at this stage propagate forward, post-hoc corrections are costly and often incomplete.

2.3 Model Training and Alignment

Training is where data becomes capability. Modern foundation models are typically developed in three phases: large-scale **pretraining** on general-purpose corpora, **fine-tuning** on curated or task-specific data, and **alignment** using methods such as Reinforcement Learning with Human Feedback (RLHF) [3].

Ethical risks here are multifaceted. Models inherit and often **amplify biases** present in their training data, manifesting as discriminatory predictions or skewed text generation [5]. **Emergent capabilities** introduce further uncertainty: systems can produce plausible misinformation, offensive content, or exploitable code without explicit instruction [9]. The **environmental footprint** of training is nontrivial; frontier LLMs require immense computational resources, contributing to carbon emissions and raising questions about the social trade-offs of scaling [11].

Alignment attempts to steer model behavior toward socially acceptable outputs, yet it is inherently normative. RLHF encodes the values of annotators and organizations, which may not reflect diverse cultural perspectives. Overreliance on alignment without addressing upstream data bias risks producing a system that *appears* safe while retaining latent ethical failures.

2.4 Environmental Accountability in AI Pipelines

Beyond ethical risks tied to bias, privacy, and misuse, modern AI pipelines carry substantial **environmental costs**. Training state-of-the-art models—particularly large neural architectures—demands extensive computation, which translates directly into energy use and associated carbon emissions [11–13]. Even well-optimized training runs can consume megawatt-hours of electricity, with additional hidden costs in data storage, cooling, and hardware manufacturing [14].

These impacts are not confined to the training stage. Data collection and preprocessing incur energy and water usage for storage, cleaning, and indexing. Inference at scale, especially for latency-sensitive applications like conversational AI, requires energy-hungry infrastructure that may operate continuously across global data centers. The environmental footprint thus extends *across* the pipeline: from initial dataset curation to deployment and post-release monitoring.

While some industry leaders have begun disclosing carbon metrics for training runs [13], systematic reporting remains rare, and standards for measurement and accountability are still evolving [12, 14]. This lack of transparency hinders both public understanding and organizational decision-making around environmental trade-offs. Without explicit budgetary and policy constraints, the drive for larger models risks locking AI development into an unsustainable trajectory.

Embedding **environmental accountability** into AI pipelines—through stage-specific targets for CO_2e, water usage, and energy efficiency—ensures that sustainability is treated as a first-class governance priority. As later sections discuss, the proposed *Eco gate* operationalizes this principle by making environmental compliance as non-negotiable as fairness, accuracy, and legal obligations.

2.5 Deployment and Inference

Once models leave the laboratory, they enter the sociotechnical domain where ethical impacts are most visible. Deployment exposes models to real users through APIs, embedded applications, or search and recommendation pipelines. At this stage, four key risks dominate:

- **Hallucination and misinformation,** where LLMs generate coherent but false outputs that can mislead users or erode public trust [6];
- **Misuse and dual-use,** as open-access generative models are repurposed for disinformation campaigns, social engineering, or automated cyberattacks;
- **Opacity and overreliance,** since users may not understand the limitations of a model's reasoning process and thus defer to it uncritically;
- **Downstream societal effects,** such as the shaping of public discourse, knowledge access, and decision-making in IR contexts.

The next generation of agentic or tool-using AI systems will expand this risk surface. When models are empowered to take autonomous actions, ethical lapses may result in immediate and tangible harm.

2.6 Post-Deployment Monitoring

A future-proof ethical framework recognizes that obligations do not end at release. Models interact with evolving environments, and both technical performance and social impact can drift over time [7]. Effective monitoring includes:

- **Auditing and logging** to capture anomalous behavior and provide forensic accountability;
- **Bias and harm tracking** to measure real-world disparities across demographic or linguistic groups;
- **Responsive remediation** to update, retrain, or even suspend systems when significant risks emerge.

Today, most organizations remain reactive rather than proactive in this stage, investing heavily in model release but underinvesting in lifecycle governance. As AI becomes more persistent and autonomous, continuous monitoring and rapid response pipelines will become non-negotiable components of ethical practice.

2.7 Conceptual Implications for NLP and IR

Framing ethics in a pipeline context has direct implications for NLP and IR. Retrieval-Augmented Generation (RAG) systems combine corpus retrieval with LLM generation, inheriting risks from both components: unvetted data sources may lead to misleading citations, while hallucination undermines factual reliability [1]. Conversational agents illustrate the tension between user engagement, transparency, and safety. Search and recommendation engines exemplify how seemingly neutral ranking choices can shape collective knowledge and public opinion.

By reconceptualizing AI ethics as a lifecycle problem, this pipeline-based perspective lays the foundation for the philosophy-to-practice bridge that the remainder of this paper develops.

3 Philosophical Foundations of Ethical AI

A robust ethical AI framework must rest on principles that are both historically grounded and actionable in contemporary sociotechnical contexts. Philosophical traditions provide enduring lenses for evaluating moral responsibility, while recent AI ethics research interprets these traditions in the context of model design, deployment, and governance [15, 16].

3.1 Consequentialism and the Logic of Outcomes

Consequentialism, most famously articulated by Jeremy Bentham and John Stuart Mill [17], judges the morality of actions by their outcomes. In AI pipelines, this translates to a focus on the *real-world impact* of models, regardless of developer intentions. Harm reduction frameworks in AI ethics are inherently consequentialist: the objective is to anticipate and minimize risks such as misinformation, systemic bias, or security misuse [5].

Contemporary applications include algorithmic impact assessments, which explicitly model the potential societal outcomes of AI deployments [18]. In NLP and IR, this perspective motivates evaluation protocols that measure not only accuracy but also *downstream social effects*, such as the spread of harmful stereotypes or unequal access to information.

3.2 Deontology and the Primacy of Duties

Deontology, associated with Immanuel Kant [19], emphasizes duties and rules rather than outcomes. From this view, an AI system is unethical if it violates intrinsic moral duties—such as respecting individual privacy or avoiding deception—even if no observable harm results.

Modern AI ethics frameworks reflect this in the form of **compliance-oriented guidelines** and **rights-based approaches**, including the European Union's AI Act and the OECD AI Principles [20, 21]. Deontological reasoning underpins requirements for informed consent in dataset collection, truthfulness in system outputs, and non-discrimination in decision-support applications.

In the context of NLP and IR pipelines, deontological principles support practices like *data minimization* (collect only what is justified), *truth-preserving summarization*, and *transparent citation of sources*, even when shortcuts might improve performance metrics.

3.3 Virtue Ethics and Responsible Practice

Virtue ethics, traced to Aristotle's *Nicomachean Ethics* [22], shifts the moral lens from actions or outcomes to the character and practices of moral agents. In the AI pipeline, the "agents" are the developers, researchers, and organizations whose choices shape

system behavior. This perspective is increasingly influential in AI ethics under the banner of *responsible AI* and *ethics by design* [8, 23].

Virtue ethics motivates long-term stewardship and reflective practice: engineers are called to cultivate traits like honesty (resisting deceptive model claims), humility (acknowledging uncertainty in capabilities), and prudence (foreseeing emergent risks). In practical terms, this perspective informs governance measures like:

- Establishing cross-disciplinary ethics boards that sustain accountability beyond compliance;
- Investing in ongoing education for AI practitioners to recognize sociotechnical risks;
- Prioritizing documentation practices such as *model cards* and *datasheets* that embed reflection into engineering workflows [7, 24].

3.4 Bridging Philosophy and Pipeline Stages

While each tradition offers a distinct lens, they are complementary when mapped onto the AI pipeline described in Section 2. Consequentialism aligns naturally with post-deployment monitoring and auditing, where real-world impact is measurable. Deontology is most salient during data collection and model alignment, where duties of fairness, transparency, and privacy can be operationalized as explicit constraints. Virtue ethics provides a throughline across all stages, emphasizing the cultivation of responsible research cultures that resist the "move fast and break things" ethos in frontier AI.

This synthesis lays the foundation for an *ethics-by-design* approach: pipeline interventions are not ad hoc patches but principled actions embedded from inception to monitoring, informed by enduring moral philosophy and reinforced by contemporary AI ethics scholarship. Figure 2 visualizes this mapping between ethical traditions and AI pipeline stages as a two-dimensional matrix. The rows represent the three philosophical lenses—Consequentialism, Deontology, and Virtue Ethics,while the columns follow the four lifecycle stages from data collection to post-deployment monitoring. Each cell lists concrete engineering interventions that operationalize the corresponding philosophical principle at that stage, highlighting where different moral frameworks converge or diverge in practice.

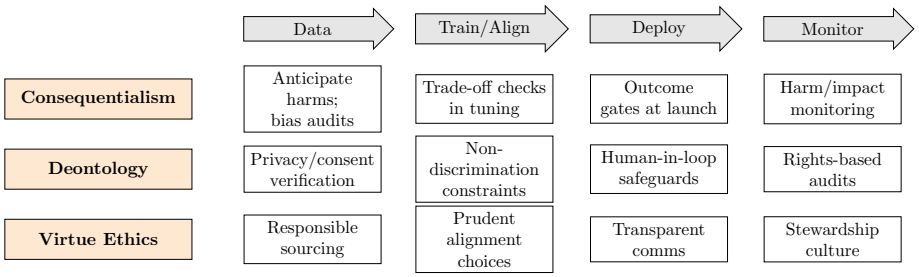


Fig. 2 Philosophy–practice matrix with compact spacing and larger font size.

4 Existing AI Ethics Frameworks and Gaps

A variety of AI ethics frameworks and governance instruments have emerged in recent years, each reflecting different regulatory philosophies and levels of enforceability. Understanding their background is essential before evaluating their gaps in addressing full AI pipelines.

4.1 Research-Led Ethical AI Pipeline Frameworks

While much attention has focused on governance guidance from regulatory bodies and standards organizations, a parallel body of scholarly work has proposed ethics-by-design frameworks emerging in the post-pandemic era. These models tend to emphasise operational integration and contextual adaptability rather than solely high-level principles.

For example, the *Hourglass Model of Organizational AI Governance* [25] introduces a multi-layered structure linking organizational policies with system-level design and lifecycle operations, enabling governance requirements to cascade through each stage of development. In the research context, the *ETHICAL* framework [26] provides actionable prompts—such as "Examine policies" and "Think about social impacts"—to guide responsible use of generative AI in academic workflows. Complementing these, Al Harbi et al. present *Responsible Design Patterns for Machine Learning Pipelines* [27], a library of reusable design components that embed ethical checks directly into technical workflows, facilitating adoption across domains.

These academic frameworks share common ground with regulatory approaches but tend to foreground researcher agency, domain-specific tailoring, and iterative ethics integration. The triple-gate model developed in this paper draws on such scholarship while extending it with an explicit environmental *Eco gate* to ensure sustainability constraints are embedded alongside fairness, accuracy, and governance checks.

4.2 Global Policy and Regulatory Frameworks

4.2.1 European Union AI Act.

The EU AI Act (2021) [20] represents the first comprehensive attempt at risk-based regulation for AI. It classifies AI systems into four tiers:

- **Unacceptable risk:** Prohibited applications such as social scoring and manipulative AI for vulnerable populations.
- **High risk:** Requires conformity assessments, human oversight, and detailed documentation (e.g., medical AI, recruitment tools).
- **Limited risk:** Must provide user disclosures when AI is in use (e.g., chatbots, deepfakes).
- **Minimal risk:** Permitted with no specific obligations (e.g., AI-powered spam filters).

Its pipeline implications are clearest at deployment and post-deployment stages, where documentation and human oversight are mandated. However, the Act is less prescriptive about upstream practices such as data curation and bias mitigation, leaving room for interpretation and variability across member states.

4.2.2 NIST AI Risk Management Framework.

The NIST AI RMF 1.0 (2023) [28] reflects a U.S. approach that favors voluntary standards and lifecycle integration. The framework organizes governance around four functional pillars: *map, measure, manage, and govern.* Unlike the EU's regulatory stance, NIST RMF directly references model development pipelines, encouraging risk awareness during data collection, training, deployment, and monitoring.

Its "govern" pillar supports organizational accountability, while "measure" and "manage" encourage proactive auditing of bias and robustness. Its non-binding nature, however, means that adoption varies widely across industry sectors.

4.3 Industry Standards and Voluntary Commitments

4.3.1 IEEE 7000 Series and Standards-Based Ethics.

The IEEE 7000 series [29] codifies engineering practices for value-based system design. IEEE 7000-2021 provides a structured process for:

1. Identifying stakeholders and their value concerns;
2. Translating values into technical system requirements;
3. Iteratively validating that system behavior aligns with declared ethical principles.

Table 1 summarizes these four frameworks across their lifecycle coverage, regulatory strength, and primary philosophical orientation. This comparison highlights their differing emphases and helps identify where pipeline stages remain under-addressed.

Table 1 Comparison of major AI ethics frameworks.

Framework	Data	Train	Deploy	Monitor	Regulatory	Lens
EU AI Act [20]	P	P	Y	Y	Binding	Duty/rights
NIST AI RMF 1.0 [28]	Y	Y	Y	Y	Voluntary	Risk/outcomes
IEEE 7000 series [29]	Y	Y	P	P	Voluntary	Practice/design
EbD-AI [30]	Y	Y	Y	Y	Voluntary	Values/lifecycle

Legend: Y = clear coverage; P = partial; – = limited/not explicit.

Compared to the EU AI Act and NIST AI RMF, the IEEE 7000 series is highly developer-centric, embedding ethical deliberation within design stages rather than relying solely on regulatory enforcement or post-hoc auditing. EbD-AI, in contrast, operationalizes a fixed set of normative values across all lifecycle stages, offering strong conceptual coverage but—like IEEE 7000—remaining voluntary. Taken together, these contrasts show that while voluntary frameworks (IEEE, EbD-AI, NIST) provide

rich design guidance, their lack of enforceability may limit their capacity to address systemic risks at scale compared to binding instruments like the EU AI Act.

4.4 Observed Gaps Across Frameworks

Despite their contributions, existing frameworks exhibit three persistent gaps:

- **Pipeline incompleteness:** Coverage is uneven across stages, with greater specificity at deployment than in upstream data governance and model pretraining.
- **Limited cross-jurisdictional alignment:** Regulatory fragmentation between regions hinders global AI governance.
- **Weak operationalization:** High-level principles often lack concrete methods for auditing emergent behaviors in frontier models.

While most frameworks exhibit some degree of lifecycle coverage, their depth and enforceability vary. For example, the NIST AI RMF explicitly addresses all stages—from data collection to post-deployment monitoring—and encourages upstream risk awareness during data curation and model training. However, its voluntary nature means these provisions lack the systemic enforcement power of the EU AI Act. Conversely, the EU AI Act offers binding obligations, particularly for high-risk systems, but is more prescriptive at deployment than at earlier pipeline stages.

Similarly, Brey and Dainow's Ethics by Design for AI (EbD-AI) framework [30] stands out for translating six core values into phase-specific engineering tasks across the AI lifecycle, embedding ethics directly into standard development methodologies. Yet, EbD-AI's preselected value set, qualitative task orientation, and voluntary adoption limit its adaptability and enforceability—especially for rapidly evolving paradigms such as multimodal or agentic AI.

These distinctions suggest that the limitation is not *awareness* of the full lifecycle, but the *operationalization* and *enforcement* of that awareness in different governance contexts. The framework proposed here aims to combine the RMF's lifecycle scope with the enforceability of risk-based regulation, while adding the concrete, metric-driven operational detail absent from both regulatory and voluntary approaches.

5 Related Work

Research on AI ethics in NLP and IR has evolved alongside the technological shifts from symbolic AI to deep learning and, most recently, foundation models. This section situates the proposed ethics-by-design framework in the context of prior work spanning three domains: (1) AI ethics principles and guidelines, (2) practical interventions for NLP and IR pipelines, and (3) emerging studies on ethics in large-scale and multimodal AI.

5.1 AI Ethics Principles and Guidelines

A significant body of work has attempted to codify the ethical responsibilities of AI systems and their developers. Early contributions include high-level principles such

as fairness, accountability, transparency, and privacy (FAT/ML) [8, 16]. These efforts informed global policy initiatives, including the OECD AI Principles [21], the EU AI Act [20], and the UNESCO Recommendation on AI Ethics (2021). While these guidelines established a normative baseline, scholars have noted that principles alone cannot guarantee ethical AI in practice [15]. The "principles-to-practice gap" persists because such frameworks rarely provide operational detail for data curation, model alignment, or post-deployment monitoring.

5.2 Pipeline-Oriented Interventions

As NLP and IR systems grew in scale and societal impact, researchers proposed interventions to embed ethics directly into model pipelines. Model cards [7] and datasheets for datasets [24] introduced structured documentation to increase transparency and support downstream auditing. Algorithmic impact assessments (AIA) [18] and red-teaming practices [9] aimed to anticipate harmful outcomes during and after deployment.

In IR contexts, studies examined bias-aware document ranking, fairness in recommender systems, and mitigation of hallucination in LLM-based retrieval-augmented generation (RAG) [1, 6]. Despite these contributions, most interventions are reactive or isolated to single stages of the pipeline, highlighting the need for a unified lifecycle framework.

5.3 Ethics in Large-Scale and Frontier AI

Recent literature recognizes that foundation models and LLMs pose unprecedented ethical challenges due to their emergent capabilities, opacity, and potential for dual use [2, 5]. Researchers documented systemic biases, privacy risks, and environmental costs of training [4, 11], and called for multidisciplinary oversight that spans the entire AI lifecycle. Initiatives like NIST AI RMF [28] and IEEE 7000 series [29] reflect this lifecycle approach but rely heavily on voluntary adoption.

Recent work also highlights gaps in cross-jurisdictional governance and the difficulty of auditing frontier models at scale. This motivates a shift toward proactive, pipeline-integrated ethics that combines philosophical rigor with engineering feasibility—a need addressed by the proposed framework.

5.4 Ethics by Design for AI (EbD-AI)

Brey and Dainow's "Ethics by Design for AI" (EbD-AI) methodology, developed under the EU-funded SHERPA and SIENNA projects, operationalizes six core values—human agency, privacy and data governance, fairness, well-being, transparency, and accountability—into concrete design requirements and phase-specific tasks spanning specification, design, data preparation, development, and testing [30]. Unlike principle-based guidelines such as the EU AI Act or NIST AI RMF, EbD-AI embeds ethics directly into standard development methodologies via task mapping, making ethical considerations part of everyday engineering practice rather than post-hoc audits. While its lifecycle integration is a notable strength, EbD-AI preselects a fixed

set of values, focuses on qualitative tasks rather than metricized decision thresholds, and—like other voluntary frameworks—depends on organizational uptake for enforcement.

5.5 Operational Distinctions from Existing Frameworks

Table 2 Operational differences between major frameworks and the proposed approach.

Feature	**EU AI Act**	**NIST AI RMF**	**EbD-AI**	**Proposed Framework**
Philosophy integration	Implicit rights-based	Implicit risk/outcomes	Six fixed ethics values mapped to lifecycle	Maps 3 traditions to each lifecycle stage
Trigger thresholds	Principles only	Encouraged, unspecified	Value-linked; no quantitative triggers	Metric-driven triggers per stage
Adaptability to new paradigms	Amendments required	Flexible, informal	Generic model mappable to any method; fixed values	Built-in reassessment triggers
Implementation granularity	Regulatory obligations	Voluntary best practices	Task-based interventions in dev process	Stage-by-stage engineering interventions
Enforcement	Binding for high-risk	Voluntary	Voluntary; org-dependent	Works in voluntary and binding compliance

Table 2 compares the operational characteristics of the EU AI Act, NIST AI RMF, EbD-AI, and the proposed framework. Three patterns stand out. First, existing frameworks vary in how explicitly they integrate ethical philosophy—rights-based in the EU AI Act, risk/outcomes-focused in NIST, fixed values in EbD-AI—whereas the proposed framework maps three philosophical traditions directly to each lifecycle stage. Second, only the proposed framework defines quantitative, metric-driven triggers for interventions, enabling automated escalation in CI/CD or MLOps contexts. Third, while all others require significant adaptation to address paradigm shifts, the proposed framework embeds reassessment triggers by design, allowing it to operate effectively in both voluntary and regulatory compliance environments.

The framework differs from existing approaches in three operational respects:

- **Lifecycle Integration with Philosophy.** Existing frameworks rarely embed normative reasoning directly into engineering checkpoints. The framework links consequentialism, deontology, and virtue ethics to measurable actions at each stage, ensuring moral reasoning remains operational rather than aspirational.

- **Metric-Driven Triggers.** Most guidelines provide high-level principles but lack concrete thresholds for intervention. The author prescribes tool-agnostic metrics (e.g., fairness differential > 0.1, KL divergence > 0.15) and explicit escalation triggers that can be implemented in CI/CD or MLOps pipelines.
- **Future-Proofing via Adaptability.** While other frameworks assume current paradigms, this one is designed to accommodate emerging risk classes in multimodal, agentic, and neuro-symbolic AI without wholesale redesign, using "reassessment triggers" baked into governance processes.

6 Toward a Future-Proof Ethics-by-Design Framework

Building on the surveyed philosophical traditions and existing AI governance frameworks, the author proposes a forward-looking *ethics-by-design* framework. This framework integrates normative principles into the AI pipeline from inception to post-deployment, aiming to bridge the persistent gap between high-level guidelines and engineering practice. Unlike reactive approaches that rely on post-hoc mitigation, ethics-by-design embeds moral reasoning, risk assessment, and accountability mechanisms as first-class elements of model development.

This proactive stance contrasts with the common "retrofit" model, where AI components are bolted onto pipelines designed without them. Retrofitted systems often lack the data structures, governance checkpoints, and feedback loops necessary for safe and effective AI operation, resulting in brittle integrations and higher ethical risk. In contrast, AI-first pipelines can align domain workflows, data governance, and model interfaces from inception, making it easier to apply the proposed ethics-by-design framework consistently across the lifecycle.

6.1 Philosophical Lenses

The framework integrates three traditions in moral philosophy—consequentialism, deontology, and virtue ethics—into the lifecycle of AI systems. Each is not treated as an abstract anchor but as a practical driver for engineering decisions from the outset.

Consequentialism evaluates actions by their outcomes, making it essential to anticipate, measure, and minimize both intended and unintended consequences of AI systems. In engineering terms, this translates into rigorous post-deployment harm audits, continuous monitoring for disparate impact, and escalation protocols when performance drifts toward harmful behaviour.

Deontology focuses on duties and adherence to rules regardless of the consequences. Within the AI pipeline, this becomes enforceable compliance gates during deployment, hard constraints in model alignment, and mandatory human-in-the-loop reviews for decisions in high-stakes contexts.

Virtue ethics emphasizes the cultivation of good judgement and moral responsibility among practitioners. Operationally, this drives a culture of cautious iteration, stakeholder engagement in design reviews, and institutional norms that reward raising ethical risks even when doing so slows delivery.

By linking each philosophical lens to concrete engineering practices, the framework ensures that ethical reasoning and technical implementation remain connected from the earliest stages of design.

Each philosophical lens constrains a distinct failure mode rather than serving as post hoc rhetoric: consequentialism bounds outcome harms via measurable risk thresholds, deontology enforces rights and procedural duties through non-negotiable prohibitions and documentation requirements, and virtue ethics shapes organizational character through oversight cadence, escalation norms, and accountability roles.

6.2 Core Principles

The framework rests on three mutually reinforcing principles:

- **Pipeline Integration** — Ethical oversight is mapped explicitly onto data collection, model training, deployment, and post-deployment monitoring stages. Each stage has dedicated checks for privacy, fairness, and downstream societal impact.
- **Multi-Layered Accountability** — Responsibility is distributed across individual researchers, project teams, and organizations. This reflects virtue ethics' focus on responsible practice, deontology's attention to duties, and consequentialism's emphasis on outcomes. The framework is designed for interoperability with varying legal regimes and adaptable to non-Western governance models, enabling consistent ethical oversight across jurisdictions with differing regulatory maturity and cultural norms.
- **Dynamic Adaptation** — As AI capabilities evolve—toward multimodal, agentic, or neuro-symbolic systems—the framework supports iterative reassessment of risks and principles, ensuring that oversight keeps pace with technological change. This includes reliance on tool-agnostic metrics such as fairness differentials, energy use per training hour, or embedding drift scores, and evaluation methods like bias audit pipelines, adversarial red-team simulations, and post-deployment harm incidence tracking to maintain consistent comparability across architectures and modalities.

In practical terms, "future-proof" in this context means three things: (i) the inclusion of explicit triggers for reassessment when model capabilities, usage contexts, or societal expectations change; (ii) reliance on tool-agnostic metrics—such as fairness differentials, energy use per training hour, or embedding drift scores—that remain applicable across architectures and modalities; and (iii) a governance process designed to incorporate unanticipated risk classes without requiring wholesale redesign of the framework. By treating adaptability as an operational requirement rather than an aspirational quality, the framework seeks to remain relevant through successive AI paradigm shifts.

6.3 Anticipating Risks in Emerging AI Paradigms

The framework's future-proofing claim rests on its adaptability to paradigms whose risk profiles are not yet fully known. To illustrate, consider three examples:

- **Multimodal Systems:** A cross-modal model may infer sensitive attributes (e.g., health status from voice, socio-economic class from home interior) without explicit user disclosure, raising novel consent and profiling concerns.
- **Agentic AI:** Autonomous decision-making agents coordinating across networks may trigger cascading effects—such as self-initiated transactions or negotiations—that lack human-in-the-loop oversight at critical junctures.
- **Neuro-symbolic Models:** Symbolic reasoning components can produce discrete, unambiguous outputs that appear authoritative but embed unexamined logical biases; their opacity may differ from deep nets but still hinder auditability.

These scenarios illustrate that "future-proof" cannot mean "unchanging"; rather, it implies the framework's capacity for iterative reassessment as new risk classes emerge.

6.4 Methodological Commitment and Preregistration

This paper contributes a design specification and evaluation protocol rather than empirical results from executed experiments. This separation is intentional and methodologically motivated. Empirical evaluation of gate-based ethics frameworks faces a fundamental challenge: threshold selection and metric calibration are inherently domain- and context-dependent, and post-hoc adjustment of thresholds after observing outcomes introduces selection bias and undermines scientific rigor.

To address this, the framework is accompanied by a preregistered evaluation protocol (Appendix A) that specifies endpoints, sampling procedures, labeling criteria, and analysis plans *before* empirical execution. This preregistration serves multiple purposes: (i) it ensures falsifiability by defining success criteria ex ante, (ii) it prevents post-hoc threshold tuning that could inflate apparent effectiveness, (iii) it enables reproducible evaluation across different organizations and domains, and (iv) it makes explicit the trade-offs between false-positive rates (blocking safe deployments) and false-negative rates (allowing risky deployments) that any gate-based system must navigate.

The protocol specifies primary endpoints (gate false-positive and false-negative rates by stage and metric), secondary endpoints (risk reduction, time-to-detection, emissions–quality–latency trade-offs), and sensitivity analyses across predefined threshold bands. By committing to these specifications before execution, the framework can be evaluated without the circularity that would arise from tuning thresholds to match observed outcomes. This methodological commitment positions the work as a falsifiable design contribution rather than a post-hoc case study.

6.5 Operationalizing the Framework

This section first presents an end-to-end illustration of the triple-gate across the lifecycle, then details a reference implementation (Section 6.5.6), and finally points to the preregistered evaluation protocol (Appendix A).

To translate principles into actionable engineering steps, the four-stage operational loop is expanded with concrete metrics, tools, and illustrative examples. Each stage

is treated as a continuous practice rather than a one-time checkpoint, ensuring ethics remains embedded throughout the AI lifecycle.

Although presented sequentially for clarity, these stages form a cyclical process rather than a one-way pipeline. Insights from post-deployment monitoring feed directly into updates in data governance and alignment practices; alignment decisions may prompt changes to deployment protocols; and new risks identified at any point can trigger a full reassessment starting from the data stage. This feedback loop ensures that the framework is capable of continuous self-correction as contexts and capabilities evolve.

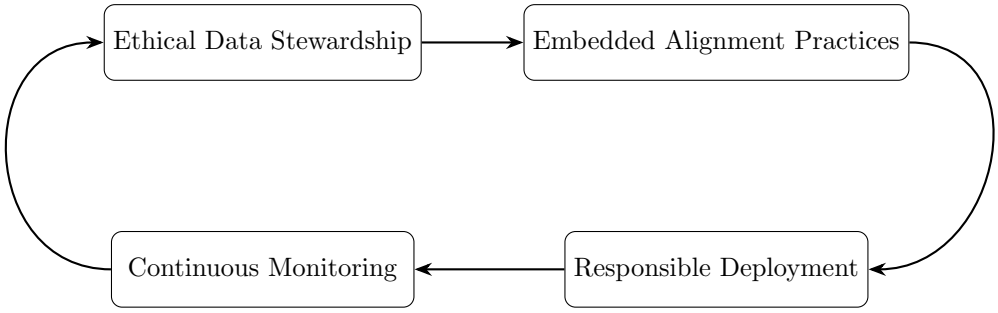


Fig. 3 Cyclical structure of the operational loop, where each stage can inform and trigger updates to earlier stages.

6.5.1 Stage 1: Ethical Data Stewardship.

Before training begins, datasets undergo structured assessment using *datasheets for datasets* [24] and algorithmic impact templates [18]. Practitioners evaluate coverage (e.g., data coverage ratio, language representation index), verify consent, and detect over- or under-represented communities; lightweight profiling tools such as `pandas-profiling` [31] and `Great Expectations` [32] can automate much of this work. In parallel, sourcing and curation decisions incorporate sustainability as a first-class constraint: minimize redundant data to reduce downstream compute, prefer storage and processing in low-carbon cloud regions, and log estimated storage/transfer energy and cooling water use so that environmental cost is visible at the same decision points as fairness and quality [10, 11]. For example, a multilingual NLP project might quantify the proportion of documents per ISO 639-1 language code and flag those falling below a 2% threshold for augmentation *while* recording the projected storage and transfer footprint, enabling a trade-off analysis between representational coverage and carbon/water budget before data growth locks in higher lifecycle costs. We use ISO 639-1 language codes as a non-PII proxy group suitable for multilingual support, and—where lawful and appropriate—layer additional groups to capture intersectional effects under privacy-preserving aggregation.

6.5.2 Stage 2: Embedded Alignment Practices.

During fine-tuning and alignment, ethical checkpoints are woven into standard evaluation workflows. Bias is measured across protected attributes using metrics such as *demographic parity difference* (DPD) and *equalized odds* (EO), with toolkits like IBM `AIF360` [33] or Microsoft `Fairlearn` [34]. In parallel, environmental performance is monitored with tools such as `CarbonTracker` [35], CodeCarbon, or ML CO_2 Impact, which log electricity consumption, carbon emissions (kg CO_2e), and, where available, water usage during each training and evaluation cycle [10, 11]. This ensures that model quality metrics (e.g., accuracy, fairness) are contextualized with sustainability costs, enabling trade-offs such as early stopping or low-precision training when marginal accuracy gains carry disproportionate environmental impact. For example, when aligning an LLM with RLHF, annotator pools should be selected for value diversity, bias reports generated after each epoch to guide dataset rebalancing, and training schedules adjusted to run in regions/times with higher renewable energy availability to reduce carbon intensity.

Worked example: Consider a multilingual customer support chatbot deployed in both the U.S. and the EU. In the U.S., *protected attributes* typically include race, gender, religion, and disability status under U.S. Equal Employment Opportunity Commission (EEOC) guidelines; in the EU, the General Data Protection Regulation (GDPR)'s "special categories" extend protection to attributes such as political opinions, trade union membership, and biometric identifiers. Suppose the model recommends a refund in 65% of cases for Group A and 55% for Group B. The demographic parity difference (DPD) is calculated as:

$$\text{DPD} = |0.65 - 0.55| = 0.10$$

If the framework threshold is set to DPD > 0.1, this result would be flagged for mitigation. Simultaneously, environmental logs might reveal that training this alignment phase consumed 450 kWh of electricity and 350 liters of cooling water, prompting consideration of smaller model checkpoints or quantization in the next cycle to meet sustainability thresholds.

6.5.3 Stage 3: Responsible Deployment and Guardrails.

Prior to release, pre-deployment audits stress-test the model against misuse scenarios and high-risk failure modes. These include adversarial red-teaming [9] to measure prompt injection success rates, automated toxicity checks via APIs such as Perspective [36], and factual verification through retrieval-based methods. Alongside safety and accuracy, deployment readiness assessments should incorporate an environmental readiness check: projecting the carbon footprint, electricity demand, and water usage of the intended inference workload [10, 11]. For high-throughput or latency-critical applications, strategies such as model distillation, quantization, or on-device inference can reduce operational energy use, while deploying on data centers powered by renewable energy mitigates climate impact.

Guardrails—such as explicit refusal templates, safe-mode fallbacks, and rate-limiting for high-cost queries—are tuned based on these audits. For example, if a

prompt injection success rate exceeds 1%, guardrail triggers might be tightened and a deployment delay introduced until mitigations are validated; likewise, if projected emissions exceed the organization's per-month CO_2e budget, model optimization or capacity throttling may be enforced before go-live.

Pre-deployment serving-emissions budgets are treated as scenario-based estimates (low, expected, high) and are re-bound at the first monitoring window; from that point, the Monitor-stage Eco gate serves as the canonical arbiter that supersedes earlier estimates.

6.5.4 Stage 4: Continuous Monitoring and Iterative Governance.

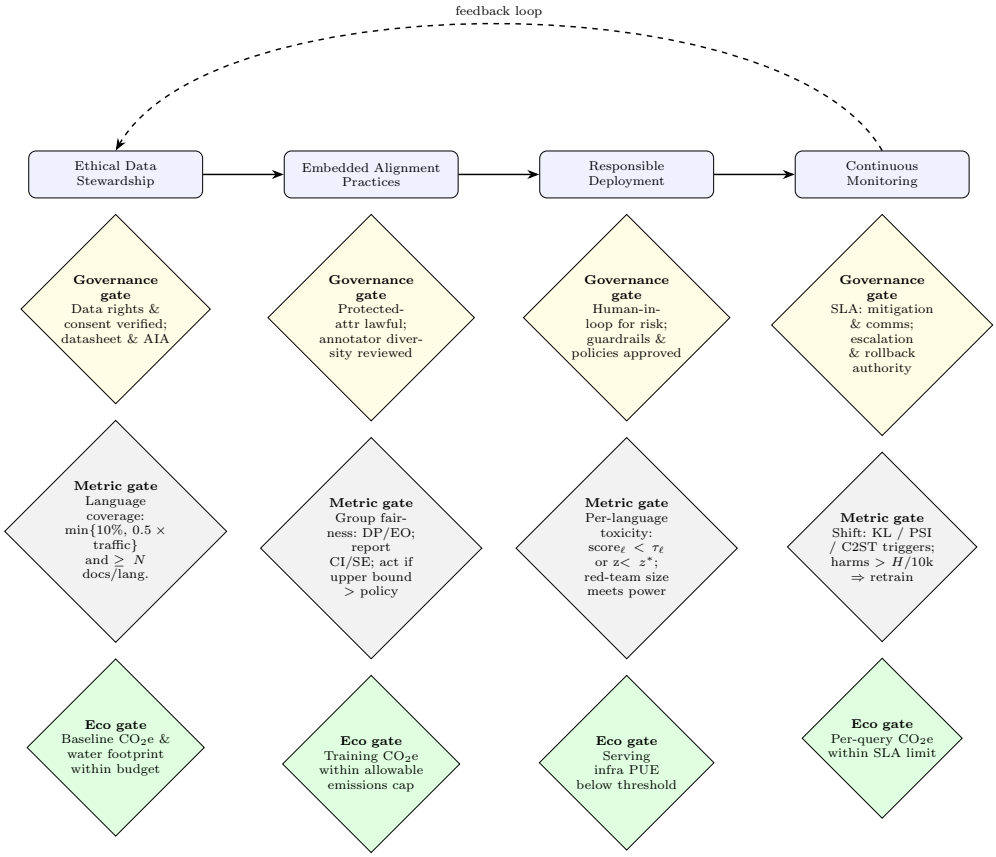
After deployment, monitoring captures both technical drift and social impact. Drift scores (e.g., KL divergence between embedding distributions over time) can signal shifts in model behaviour, while harm incidence rates and stakeholder satisfaction indices track real-world effects. Tools such as `evidently.ai` [37] can automate drift detection, while user-facing portals allow stakeholders to report harms.

Environmental performance should also be monitored as part of the governance cycle, since inference workloads at scale can have non-trivial carbon and water footprints. Following Strubell et al. [11], operators can periodically estimate CO_2-equivalent emissions by logging inference GPU-hours, local grid carbon intensity, and data centre PUE (Power Usage Effectiveness). For facilities using evaporative cooling, water usage metrics—highlighted in Crawford's analysis of AI infrastructure costs [10]—should be tracked, especially in water-scarce regions.

Worked example: A deployed multilingual chatbot serving 20 million queries per month logs its total inference GPU-hours and calculates monthly CO_2 emissions using local grid emission factors. If emissions exceed a set budget (e.g., 50 kg CO_2e per million queries), governance protocols require optimization actions—such as model distillation, quantization, or routing low-risk queries to smaller models—before the next reporting cycle. Similarly, water usage per query is monitored, and if cooling-related consumption surpasses regional conservation thresholds, workloads may be migrated to less water-stressed facilities.

6.5.5 End-to-End Illustration.

Consider a conversational AI deployed in a multilingual customer support setting. During Ethical Data Stewardship, the dataset is profiled with `pandas-profiling` to confirm that each target language accounts for at least 10% of the training corpus. In Embedded Alignment Practices, IBM `AIF360` computes a demographic parity difference for resolution suggestions by customer gender, flagging values exceeding 0.1 as requiring mitigation. Responsible Deployment includes a Perspective API toxicity check, requiring a mean score below 0.05 on a curated red-team prompt set before go-live. Finally, Continuous Monitoring with `evidently.ai` triggers retraining if KL divergence in response embeddings over a 30-day window exceeds 0.15, or if monthly user harm reports surpass a pre-set tolerance of five incidents.



Reading the diagram. Each stage has three gates in this order: *Governance* (yellow), *Metric* (gray), and *Eco* (green). Eco gates enforce stage-specific environmental limits (CO_2e, water, energy) alongside performance and compliance criteria.

Fig. 4 Triple-gate lifecycle for a multilingual support. Gates run as CI/CD jobs; failure blocks promotion; successes emit signed artefacts.

The triple-gate structure in Figure 4 illustrates how each stage of the pipeline is governed by a quantitative *Metric gate*, a qualitative *Governance gate*, and an environmental *Eco gate*. At the **Ethical Data Stewardship** stage, the Metric gate enforces a language coverage target—such as a minimum of 10% or half the observed traffic share, with an absolute document count threshold—while the Governance gate verifies data rights and consent through datasheets [24] and algorithmic impact assessments [18]. The Eco gate estimates dataset storage and preprocessing energy using tools such as `CodeCarbon` or `CarbonTracker`, reporting emissions in CO_2e and estimated water usage following emerging sustainable AI reporting standards [12, 14].

In **Embedded Alignment Practices**, the Metric gate applies group fairness tests such as demographic parity (DP) and equalized odds (EO) [38], reporting confidence intervals (CI) and standard errors (SE) to ensure statistical robustness, while governance reviews check that handling of protected attributes complies with applicable law

and that annotator pools are demographically diverse. The Eco gate logs per-epoch GPU hours, electricity source mix, and cooling water usage, enabling carbon intensity benchmarking [13].

The **Responsible Deployment** Metric gate measures per-language toxicity, requiring scores in each supported language to remain below a set threshold τ_ℓ or a critical z-score, with red-team prompt sets large enough to meet statistical power requirements [9]; its Governance gate mandates human-in-the-loop sign-off for high-risk use cases and formal approval of guardrails and policies [8]. The Eco gate conducts a final pre-release life-cycle assessment, estimating cumulative CO_2e emissions from training to inference, and flags deployments that exceed the organization's environmental budget for review [12].

Finally, **Continuous Monitoring** uses shift-detection metrics, such as Kullback–Leibler (KL) divergence [39], Population Stability Index (PSI) [40], and Classifier Two-Sample Test (C2ST) [41]—to trigger retraining when any exceed thresholds, and tracks severity-weighted user harm rates per 10,000 interactions (H), while the Governance gate enforces service-level agreements (SLA) for mitigation, communication, and escalation. The Eco gate monitors inference energy per request, aggregated CO_2e over time, and water consumption in cooling, ensuring that efficiency degradations are caught alongside accuracy drops [13, 14].

Together, the three gates at each stage ensure that measurable performance, compliance, and sustainability criteria are met before progression, reducing the likelihood of ethical, technical, or environmental harms propagating downstream.

The illustration above specifies what the triple-gate checks and why. The next subsection translates that policy logic into a concrete implementation blueprint suitable for CI/CD and runtime serving. Readers focused on mechanics and audit artefacts can continue to Section 6.5.6; those looking for the empirical protocol can jump to Appendix A.

6.5.6 Reference implementation of the triple-gate

Building on the end-to-end illustration, this subsubsection specifies how to integrate the triple-gate into CI/CD and serving, and what artefacts are produced for audits. It is implementation-oriented and does not report quantitative results.

Gate API contract. Each stage exposes a gate endpoint that consumes a metrics bundle and returns a signed decision with justifications and pointers to artefacts.

```
POST /gate/{stage}
Request:
  metrics: {metric_name: value, ...}
  metadata: {commit_id, model_id, timestamp, actor}
Response:
  decision: pass|block|throttle
  reasons: [ {metric, value, threshold, rule_id}, ... ]
  artefacts: [ {path, sha256, signer_id}, ... ]
  audit_id: UUID
```

CI/CD hook. The gates run as blocking jobs; failures prevent promotion and emit auditable logs.

```
stages: [build, test, gates, deploy]
jobs:
  - name: gate:data
    run: gates.py --stage data --config cfg.yaml --out artefacts/data/
    on_fail: block
  - name: gate:train_align
    run: gates.py --stage train_align --config cfg.yaml --out artefacts/train/
    on_fail: block
  - name: gate:deploy
    run: gates.py --stage deploy --config cfg.yaml --out artefacts/deploy/
    on_fail: block
  - name: gate:monitor
    run: gates.py --stage monitor --config cfg.yaml --out artefacts/monitor/
    on_fail: alert
```

Artefact manifest. Gate runs emit a compact manifest used for conformity assessment and audits.

```
audit_id: UUID
stage: data|train_align|deploy|monitor
entries:
  - gate: metric|governance|eco
    metric: name
    value: float|bool
    threshold: float|bool
    decision: pass|block|throttle
    evidence: path/to/file
    timestamp: ISO-8601
    signer: key-id
```

Threat-scenario walkthroughs

Three scenarios illustrate the trigger → automated action → oversight pattern. No empirical claims are made here; thresholds are defined in the configuration and may vary by domain.

Scenario A — Data fairness regression. Trigger: demographic parity difference exceeds the configured threshold for a language group. Action: block promotion and open a remediation ticket with the offending slice and seed examples. Oversight: responsible AI reviewer approves release only after a fresh gate pass.
Scenario B — Post-alignment toxicity spike. Trigger: toxicity mean breaches the stage threshold or misuse block-rate falls below the minimum. Action: block and roll back to the last green model snapshot, then re-run alignment tests. Oversight: safety lead signs off on the mitigation and re-test report.

Scenario C — Eco budget breach at deploy. Trigger: projected serving emissions per query exceeds the regional budget. Action: throttle or block; propose quantization or routing to a smaller model. Oversight: platform owner records trade-offs and signs the release record.

These walkthroughs operationalize the triple-gate as enforceable controls with auditable artefacts. This concludes the reference implementation; the preregistered empirical protocol is specified in Appendix A.

All thresholds are pre-registered (Appendix A) and sensitivity-tested within pre-defined bands, and governance-gate reviews discourage metric gaming and require documented justification for any post hoc change.

6.6 Adaptation Pathways for Cross-Context Deployment

A "future-proof" framework must not only survive paradigm shifts but also adapt across legal regimes, sectors, and cultural contexts. The author proposes three adaptation pathways (Figure 5):

1. **Regulatory Mapping:** Align each framework stage with local legal instruments (e.g., data minimization under GDPR, algorithmic transparency under the U.S. Algorithmic Accountability Act) while maintaining the same internal checkpoints.
2. **Cultural Norm Alignment:** Adjust fairness and transparency metrics to reflect local normative expectations (e.g., group fairness definitions in collectivist societies vs. individual fairness in individualist societies).
3. **Sectoral Customization:** Tailor metric thresholds and audit frequency to domain risk profiles (e.g., higher scrutiny for healthcare AI than for recommendation engines).

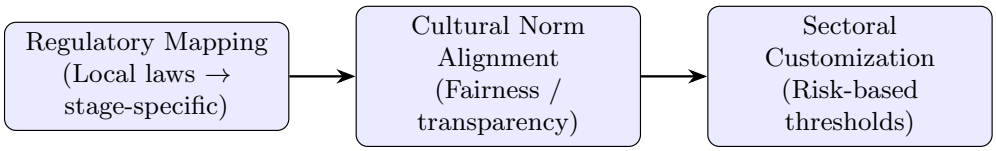


Fig. 5 Adaptation pathways for deploying the framework across jurisdictions and sectors.

These pathways allow for local compliance without sacrificing the framework's internal coherence, ensuring that adaptations remain systematic rather than ad hoc.

6.7 Regulatory crosswalk: mapping gates to EU AI Act and NIST AI RMF

As summarized in Table 3, the crosswalk is an illustrative mapping from the triple-gate controls to high-level obligations in the EU AI Act and to NIST AI RMF functions.

The intent is alignment at the level of control purpose rather than article-by-article compliance. Governance gates map to documentation, oversight, transparency, and post-market duties; Metric gates map to measurement and pre-release safety testing; Eco gates are treated as risk-management and reporting artefacts even where the Act is silent on explicit environmental thresholds. In practice, these controls run as CI/CD gate checks whose pass or fail produces auditable artefacts for conformity assessment and ongoing risk management. Entries marked as organizational policy indicate controls that extend beyond the Act's minimum while remaining compatible with NIST functions.

Table 3 Crosswalk from triple-gate controls to EU AI Act obligations and NIST AI RMF functions.

Pipeline stage	Gate	EU AI Act (shorthand)	NIST AI RMF
Data	Governance	data governance; documentation	Map, Measure, Govern
Data	Metric	bias; data quality	Measure, Manage
Data	Eco	risk management; reporting	Measure, Govern
Train/Align	Governance	human oversight; records	Govern, Manage
Train/Align	Metric	accuracy; robustness tests	Measure, Manage
Train/Align	Eco	energy and water logs (org policy)	Measure, Govern
Deploy	Governance	transparency; human-in-the-loop	Govern, Manage
Deploy	Metric	pre-release safety tests	Measure, Manage
Deploy	Eco	serving budgets (org policy)	Measure, Manage
Monitor	Governance	post-market monitoring; incidents	Govern, Manage
Monitor	Metric	drift and harm triggers	Measure, Manage
Monitor	Eco	footprint tracking (org policy)	Measure, Govern

7 Discussion

In this research, the Eco gate is positioned a first-class blocker: deployments may pass accuracy and fairness yet still be halted if projected emissions or water use exceed the organizational budget. Treating sustainability as a gating criterion, rather than a post-hoc report, rebalances incentives for model size, quantization, and placement.

Embedding gates in CI/CD and runtime serving paths is the crucial shift from "principles and documentation" to "operational control." The gate outputs decide promotion, rollback, or throttling, and their logs become audit artefacts for conformity assessment. This is where the framework substantively departs from prior lifecycle ethics work.

The distinction between an organizational framing and a control architecture is fundamental. Many ethical failures in AI systems are not failures of principle—the principles are often well-understood—but failures of *control*: the absence of enforceable checkpoints that prevent harmful configurations from progressing through the pipeline. This parallels safety-critical engineering domains: aviation safety depends not only on design principles but on control topologies (redundant systems, automated fail-safes, mandatory pre-flight checks) that enforce those principles operationally. Similarly, DevSecOps practices embed security gates into CI/CD pipelines not as documentation exercises but as blocking controls that prevent vulnerable code from reaching production.

The triple-gate framework positions ethics governance as a control topology rather than a checklist. Each gate functions as a circuit breaker: when thresholds are breached, the system blocks progression and requires remediation before resumption. This is qualitatively different from post-hoc reporting or voluntary best practices, which rely on organizational discipline rather than automated enforcement. The proposed framework differs from prior lifecycle ethics approaches in its deliberate integration of normative reasoning with concrete engineering checkpoints from project inception. While frameworks such as the NIST AI RMF or IEEE 7000 series recognize lifecycle coverage, their voluntary nature and abstract treatment of philosophical principles often leave operational details underspecified. By contrast, this approach treats philosophy not as a preamble but as an ongoing operational constraint, with metrics and governance triggers woven into each stage as enforceable controls.

This philosophy-to-practice orientation has implications for how AI is integrated into domain workflows. In the "retrofit" model—where AI components are appended to pipelines designed without them—core data structures, governance hooks, and feedback loops are often absent. This absence makes alignment retrofits brittle, increases risk, and constrains ethical guardrails. In contrast, AI-first pipelines can align domain processes, data governance, and model interfaces from inception, making it easier to apply lifecycle ethics consistently.

Adoption will face practical challenges. Organizational inertia may resist structural changes to pipelines. Costs associated with implementing continuous monitoring and bias audits may deter smaller organizations. Cross-jurisdictional applicability demands adaptation to non-Western governance norms, local regulatory regimes, and culturally specific conceptions of fairness and accountability. Addressing these challenges will require not only technical tooling but also policy incentives and sector-specific adaptation guidelines.

By embedding Eco gates throughout the pipeline, environmental impact assessment shifts from an afterthought to a first-class design requirement. This operationalizes recommendations from prior work [12–14] by making carbon and energy tracking mandatory at each stage—e.g., dataset collection, model training, serving, and post-deployment monitoring. The Eco gate enables actionable trade-off analysis: a deployment might pass accuracy and fairness tests but be halted if projected emissions exceed organizational or policy thresholds. In doing so, the framework reframes sustainability not as a constraint that slows innovation, but as a proactive criterion that can guide model architecture, hardware choice, and deployment strategy.

7.1 Limitations

Several limitations constrain the framework's applicability and should inform its adoption. First, this paper presents a design specification and preregistered evaluation protocol rather than empirical results from executed experiments. While this separation is methodologically motivated (see Section 6.4), it means that claims about effectiveness, false-positive/false-negative rates, and risk reduction remain untested until the protocol is executed. The framework's value as a control architecture is independent of empirical validation, but its practical impact depends on demonstrating that gates reduce harmful deployments without excessive false positives.

Second, threshold selection and calibration are inherently sensitive to domain, context, and organizational risk tolerance. The framework specifies that thresholds should be preregistered and sensitivity-tested, but it cannot prescribe universal values. This creates a risk of threshold gaming: organizations might set permissive thresholds to avoid blocking deployments, undermining the framework's protective function. Governance gates mitigate this by requiring documented justification for threshold choices and independent review, but the risk remains, especially in voluntary adoption contexts.

Third, the framework assumes a baseline level of organizational maturity: it requires CI/CD infrastructure, monitoring capabilities, and governance processes that may be absent in smaller organizations or early-stage projects. The costs of implementing continuous monitoring, bias audits, and environmental tracking may deter adoption, particularly for resource-constrained teams. This limits the framework's immediate applicability to organizations with established MLOps practices.

Fourth, metric selection and fairness definitions embed cultural and normative assumptions. Group fairness metrics (demographic parity, equalized odds) reflect Western individualist conceptions of fairness that may not align with collectivist cultural norms. The framework's adaptation pathways (Section 6.6) address this partially, but the underlying metric choices remain culturally situated. Similarly, protected attribute definitions vary across jurisdictions, requiring careful mapping to local legal frameworks.

Fifth, environmental impact measurement—particularly for the Eco gate—carries significant uncertainty. Carbon footprint estimation depends on grid emission factors, data center PUE values, and hardware efficiency metrics that may be unavailable or imprecise. Water usage measurement is even more uncertain, as cooling water consumption varies with climate, facility design, and operational practices. These measurement challenges mean that Eco gate decisions may be based on noisy estimates, potentially leading to both false positives (blocking low-impact deployments) and false negatives (allowing high-impact deployments).

Finally, the framework's effectiveness depends on organizational buy-in and cultural change. Technical controls can be circumvented if organizational incentives reward speed over safety, or if governance processes lack enforcement authority. The framework addresses this through virtue-ethical emphasis on responsible practice and multi-layered accountability, but it cannot guarantee adoption or prevent organizational resistance to structural changes.

8 Conclusion

This paper has proposed a *future-proof ethics-by-design framework* for NLP and IR that integrates consequentialist, deontological, and virtue-ethical reasoning into the AI pipeline from data collection to post-deployment monitoring. By bridging the principles-to-practice gap, the framework operationalizes philosophical commitments as measurable, enforceable lifecycle interventions.

The framework's contribution lies in its operational control architecture: embedding moral reasoning, risk assessment, and accountability mechanisms as first-class

elements of system architecture, rather than retrofitting ethics onto legacy pipelines. This AI-from-inception philosophy enhances resilience to paradigm shifts, from unimodal NLP to multimodal, agentic, and neuro-symbolic systems.

Future work should prioritize developing open-source tooling for continuous ethics auditing, refining metrics that remain robust across model classes, and establishing governance templates adaptable to varying legal and cultural contexts. For practitioners, the imperative is clear: ethics cannot be an afterthought; it must be an organizing principle from day one.

Declarations

- Funding: This research received no funding.
- Conflict of interest: The corresponding author states that there is no conflict of interest.
- AI-assisted writing: Large language models were used for language polishing and structural editing during manuscript preparation. They were not used for conceptual development, metric design, threshold specification, or analysis. All framework design decisions, threshold choices, and analytical content were determined by the author.
- Citation: Catapang, J.K. Building the ethical AI framework of the future: from philosophy to practice. AI Ethics 6, 150 (2026). https://doi.org/10.1007/s43681-026-01003-8

Appendix A Preregistered evaluation protocol

A.1 Objectives

1. Estimate gate false-positive and false-negative rates by stage and metric.
2. Quantify pre-release risk reduction and runtime detection performance relative to baseline.
3. Characterize emissions–quality–latency trade-offs under multiple eco budget regimes.

A.2 Endpoints

1. Primary endpoints:

$$\mathrm{FP}_{\mathrm{gate}} = \Pr(\text{block} \mid \text{safe}), \qquad \mathrm{FN}_{\mathrm{gate}} = \Pr(\text{pass} \mid \text{risky})$$

   reported per stage and metric with 95% bootstrap confidence intervals.
2. Secondary endpoints:
   - change in numeric risk metrics from baseline (same pipeline configuration without gates, or historical decisions if gates were not previously implemented) to with-gates,
   - time-to-detection for drift and harms,

- emissions per query and training footprint versus quality and latency across budget regimes.

3. Sample size justification: target $N_{\text{pre}} = 150$ and $N_{\text{run}} = 40$ provide sufficient power to detect FP/FN rates differing from null hypothesis (e.g., FP=0.05, FN=0.10) with 80% power at $\alpha = 0.05$ assuming moderate effect sizes, accounting for stratification and multiple comparisons.

A.3 Scope and systems

- Fix system under test, model family, deployment modes, and regions before data collection.
- Record these choices in an artefact manifest; document any later deviations.

A.4 Data sources and sampling

1. Pre-release changes: target $N_{\text{pre}} = 150$ items, stratified by change type (data refresh, alignment run, guardrail edit, configuration change) in proportions reflecting typical pipeline operations or equal allocation if proportions unknown. Record item_id, stage, timestamp, description, owner, source. Analyze when target reached or at fixed calendar date (see A.8).
2. Runtime incidents: target $N_{\text{run}} = 40$ items, stratified by severity (high, medium, low) in proportions reflecting incident distribution or equal allocation if unknown. Use the same metadata fields. Analyze when target reached or at fixed calendar date.
3. Test suites (fixed before gate runs):
   - data stage fairness slices,
   - train/align toxicity set,
   - deploy domain QA for hallucination,
   - monitor drift histograms.
4. Inclusion and exclusion: include items relevant to the stage with required metadata; exclude and list items with missing mandatory fields. For partial data (some metrics missing but others present), use complete-case analysis per metric; document missingness patterns.

A.5 Metrics and thresholds

- Data: demographic parity difference and data quality.
- Train/align: toxicity mean and robustness.
- Deploy: hallucination rate and misuse block-rate.
- Monitor: drift (KL) and harm incident rate.
- Governance checks: documentation completeness, lawful processing, human oversight, transparency controls, incident SLAs.
- Eco checks: training and serving emissions budgets and, where applicable, water-use constraints.
- Register exact thresholds and directions (max, min, eq) in a time-stamped configuration file prior to analysis.

A.6 Gate rules

1. A gate blocks if any configured metric breaches its threshold or a governance check fails.
2. Decisions are pass, block, or throttle.
3. Log metric values, thresholds, decision, evidence pointers, timestamps, and signer identifiers.

A.7 Labeling and adjudication

1. Two independent reviewers label each item safe or risky with rationale and confidence.
2. Compute Cohen's kappa prior to adjudication. If $\kappa < 0.6$, conduct calibration discussion and re-label; if still below threshold after calibration, exclude item from primary analysis and report separately. If $\kappa \geq 0.6$, apply majority vote; resolve ties by consulting a third independent reviewer (blinded to gate outcomes and prior labels) rather than defaulting to risky, to avoid conservative bias.
3. Keep reviewers blinded to gate outcomes.

A.8 Analysis plan

1. Estimation: compute false-positive and false-negative rates per stage and metric with 2000 bootstrap resamples and a fixed random seed.
2. Baseline comparison: where paired baseline decisions exist (same items evaluated with and without gates, or matched historical controls), use McNemar's test for error profile differences.
3. Numeric effects: summarize changes in risk metrics as means with bootstrap intervals.
4. Multiple comparisons: report per-metric intervals; provide a Benjamini–Hochberg adjusted summary in appendix tables.
5. Sensitivity: vary thresholds within pre-specified bands; for eco metrics vary grid intensity and cooling-water assumptions within declared ranges.
6. Stopping and missing data: analyze when targets ($N_{\text{pre}} = 150$, $N_{\text{run}} = 40$) are reached or at a fixed calendar date (specified before data collection begins). For missing data: use complete-case analysis per metric (exclude items missing that specific metric but include for other metrics); document missingness patterns and report counts of excluded items per metric. Omit unlabeled items (those failing inter-rater reliability threshold) from primary analysis and report separately.

A.9 Robustness and threats to validity

- Construct validity: spot-check metric computations against raw artefacts.
- External validity: results apply to sampled domains and languages; state limits.
- Governance verification: verify checks against artefact hashes.
- Known limitations: enumerate and discuss in the final report.

A.10 Reporting and artefact release

- Release configuration files, evaluation scripts, test suites or generators, gate logs or redacted summaries, and analysis notebooks required to regenerate tables.
- Include checksums and a license; exclude user-identifying content.

A.11 Deviations

- Document any deviation from the protocol, including threshold changes after outcome inspection, with justification and time stamp.

A.12 Results policy

- No results are reported in this manuscript version; executing the protocol will populate the tables referenced in the main text.

References

[1] Lewis, P., Perez, E., Piktus, A., *et al.*: Retrieval-augmented generation for knowledge-intensive nlp tasks. In: Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS) (2020)

[2] Bommasani, R., Hudson, D.A., Adeli, E., et al.: On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021)

[3] OpenAI: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)

[4] Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT) (2021)

[5] Weidinger, L., Mellor, J., Rauh, M., *et al.*: Taxonomy of risks posed by language models. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT) (2022)

[6] Ji, Z., Lee, N., Frieske, R., et al.: Survey of hallucination in natural language generation. ACM Computing Surveys (2023)

[7] Mitchell, M., Wu, S., Zaldivar, A., *et al.*: Model cards for model reporting. In: Proceedings of the 2019 ACM Conference on Fairness, Accountability, and Transparency (FAccT) (2019)

[8] Floridi, L., Cowls, J., Beltrametti, M., et al.: Ai4people—an ethical framework for a good ai society: Opportunities, risks, principles, and recommendations. Minds and Machines (2018)

[9] Ganguli, D., Hernandez, D., Lovitt, L., *et al.*: Predictability and surprise in large generative models. In: Proceedings of the 2022 ACM Conference on Fairness,

Accountability, and Transparency (2022)

[10] Crawford, K.: The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. Yale University Press, New Haven (2021)

[11] Strubell, E., Ganesh, A., McCallum, A.: Energy and policy considerations for modern deep learning research. In: Proceedings of the AAAI Conference on Artificial Intelligence (2020)

[12] Schwartz, R., Dodge, J., Smith, N.A., Etzioni, O.: Green ai. Communications of the ACM (2020)

[13] Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L., Rothchild, D., So, D., Texier, M., Dean, J.: Carbon emissions and large neural network training. arXiv preprint arXiv:2104.10350 (2021)

[14] Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., Pineau, J.: Towards the systematic reporting of the energy and carbon footprints of machine learning. In: Proceedings of the AAAI Conference on Artificial Intelligence (2020)

[15] Mittelstadt, B.: Principles alone cannot guarantee ethical ai. Nature Machine Intelligence (2019)

[16] Jobin, A., Ienca, M., Vayena, E.: The global landscape of ai ethics guidelines. Nature Machine Intelligence (2019)

[17] Mill, J.S.: Utilitarianism. Parker, Son, and Bourn, London (1863)

[18] Reisman, D., Schultz, J., Crawford, K., Whittaker, M.: Algorithmic impact assessments: A practical framework for public agency accountability. Technical report, AI Now Institute, New York (2018)

[19] Kant, I.: Groundwork of the Metaphysics of Morals. Johann Friedrich Hartknoch, Riga (1785)

[20] European Commission: Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts (2021)

[21] OECD: OECD Principles on Artificial Intelligence. OECD Publishing (2019)

[22] Aristotle: Nicomachean Ethics. Oxford University Press, Oxford (1953). Originally written ca. 350 BCE. Translated by W. D. Ross

[23] Whittlestone, J., Nyrup, R., Alexandrova, A., Cave, S.: The role and limits of principles in ai ethics. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES) (2019)

[24] Gebru, T., Morgenstern, J., Vecchione, B., et al.: Datasheets for datasets.

Communications of the ACM (2021)

[25] Mäntymäki, M., Minkkinen, M., Birkstedt, T., Viljanen, M.: Putting ai ethics into practice: The hourglass model of organizational ai governance. arXiv preprint arXiv:2206.00335 (2022)

[26] Eacersall, D., Pretorius, L., Smirnov, I., et al.: Navigating ethical challenges in generative ai-enhanced research: The ethical framework for responsible generative ai use. arXiv preprint arXiv:2501.09021 (2024)

[27] Al Harbi, S.H., Nganyewou Tidjon, L., Khomh, F.: Responsible design patterns for machine learning pipelines. arXiv preprint arXiv:2306.01788 (2023)

[28] Tabassi, E.: Artificial Intelligence Risk Management Framework (AI RMF 1.0) (2023)

[29] IEEE Standards Association: IEEE 7000-2021: Model Process for Addressing Ethical Concerns During System Design. IEEE, New York (2021)

[30] Brey, P., Dainow, B.: Ethics by design for ai: An approach to the development of ethically aligned ai. AI and Ethics (2024) https://doi.org/10.1007/s43681-023-00383-5

[31] Brugman, S., contributors: pandas-profiling: Generate Profile Reports from a pandas DataFrame. https://github.com/ydataai/pandas-profiling (2024)

[32] Great Expectations Team: Great Expectations: Always Know What to Expect from Your Data. https://greatexpectations.io/ (2024)

[33] Bellamy, R.K., Dey, K., Hind, M., et al.: Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development (2019)

[34] Microsoft Research: Fairlearn: Assessing and Improving Fairness in AI Systems. https://fairlearn.org/ (2024)

[35] Anthony, L.F.W., Kanding, B., Selvan, R.: Carbontracker: Tracking and predicting the carbon footprint of training deep learning models. arXiv preprint arXiv:2007.03051 (2020)

[36] Google Jigsaw: Perspective API: Using Machine Learning to Improve Conversations Online. https://perspectiveapi.com/ (2024)

[37] Evidently AI Team: Evidently AI: Evaluate and Monitor Machine Learning Models in Production. https://www.evidentlyai.com/ (2024)

[38] Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: Advances in Neural Information Processing Systems (NeurIPS) (2016)

[39] Kullback, S., Leibler, R.A.: On information and sufficiency. Annals of Mathematical Statistics (1951)

[40] SAS Help Center: Concepts: Performance Monitoring. https://documentation.sas.com/doc/en/mdlmgrcdc/v_057/mdlmgrug/p1c6xm7tthdajkn1t6esm4n3kwnq.htm (2025)

[41] Lopez-Paz, D., Oquab, M.: Revisiting classifier two-sample tests. arXiv preprint arXiv:1610.06545 (2016)