

# Responsible Machine Learning Deployment: Imperative Framework for Ethical Action

MAIKEL LEON

Department of Business Technology  
University of Miami  
Miami, Florida, USA

*Abstract:* The rapid expansion of Machine Learning (ML) across finance, healthcare, education, and public policy makes ethical oversight an imperative rather than an optional add-on. This paper responds to that urgency by proposing a comprehensive framework grounded in ten principles—accuracy, fairness, accessibility, security, privacy, transparency, accountability, human oversight, sustainability, and harm avoidance—and positioning them within existing international guidelines. Recent scoping reviews have highlighted the lack of consistent evaluation frameworks across domains and have called for systematic approaches to fairness, accountability, transparency, and ethics. Motivated by case studies of algorithmic redlining, dataset bias, hallucinations in large language models, and ecological concerns, we develop a weighted scoring rubric with thresholds to diagnose ethical compliance. We demonstrate the rubric through case studies, illustrating how the scores identify deficiencies and guide mitigation. The proposed framework is built upon the EU AI Act, NIST’s AI Risk Management Framework, UNESCO’s recommendations, and the OECD AI Principles. We reflect on AI’s energy footprint and the so-called “nuclear dependence” argument, and conclude with a roadmap for practitioners.

*Key-Words:* Machine learning, ethical computing principles, fairness, bias mitigation, accessibility, sustainability, energy and water consumption, scoring rubric

Received: April 11, 2025. Revised: June 29, 2025. Accepted: August 9, 2025. Published: January 8, 2026.

## 1 Introduction

Machine Learning (ML) is transforming nearly every facet of society, from credit scoring and medical diagnosis to autonomous vehicles and public policy. With this power comes significant ethical responsibility. Reports of mortgage approval algorithms disadvantaging applicants of color illustrate how algorithmic decision systems can replicate historical discrimination. Similarly, widely used medical datasets, such as the Pima Indians diabetes database, contain imbalanced classes and limited demographic diversity, raising concerns about fairness and representation. Large language models (LLMs) trained on web-scale data occasionally invent facts or citations, a phenomenon described as “hallucinations,” which can potentially mislead clinicians and researchers, [1]. Such misuse of data, design oversights, or opaque model behavior can lead to ethical breaches that have real social consequences.

Recent scholarship highlights the importance of developing robust, operationalizable ethical frameworks. A comprehensive scoping review of fairness, accountability, transparency, and ethics (FATE) in social media and healthcare observes that existing guidelines are heterogeneous

and inconsistent across domains, which hampers standardized evaluation, [2]. In medicine, Chen and colleagues document how algorithmic biases arise from data acquisition, genetic variation, and labeling choices, and warn that ignoring protected attributes does not eliminate unfairness; underrepresented groups can be underdiagnosed, and unequal access to care may persist even when sensitive variables are removed, [3]. Beyond healthcare, software practitioners often treat fairness as a second-class quality; they lack tools and processes to engineer fairness throughout the ML lifecycle and call for fairness-aware MLOps practices, [4]. These studies underscore the need for comprehensive, actionable principles that go beyond high-level statements.

Emerging privacy and sustainability challenges further raise the stakes. Membership inference attacks enable adversaries to determine whether a specific record was utilized in training a model, thereby exposing sensitive personal data. To mitigate such attacks, [5] demonstrates the importance of privacy-by-design and continuous monitoring by categorizing tokens for learning and unlearning during training. Environmental research highlights the material footprint of AI.

Other studies, [6], [7], apply a digital planetary

health lens to generative AI, noting that LLMs increase electronic waste, greenhouse gas emissions, and water use, and calls for multispecies justice and intergenerational considerations. It's estimated that global AI water withdrawal could reach 4.2–6.6 billion cubic meters by 2027, revealing a hidden water footprint that complements carbon emissions. These insights highlight the intertwined human, social, and planetary impacts of AI development, underscoring the importance of sustainability in ethical analysis, [8].

This paper builds upon these findings by not only synthesizing ten ethical principles but also proposing additional principles and case studies that go beyond conventional guidelines. We emphasize continuous monitoring, environmental stewardship, privacy-preserving training, and context-sensitive fairness. By doing so, we aim to provide practitioners and policymakers with a holistic, actionable framework for the ethical deployment of ML.

This paper addresses three goals. First, we synthesize ten ethical principles for the responsible deployment of ML. The principles draw on regulatory and scholarly sources, including the EU AI Act's risk-based obligations, [9], NIST's trustworthiness characteristics, [10], UNESCO's human-rights-centred recommendations, [11], the OECD AI Principles, [12], and the classical Ten Commandments of Computer Ethics, [13]. Second, we propose a weighted scoring rubric and an algorithmic process for evaluating ML systems against these principles and illustrate its use in a hypothetical mortgage-underwriting scenario. Finally, we compare our framework with existing international guidelines, evaluate ecological aspects (including AI's energy demands and the "nuclear dependence" argument), and outline a roadmap for responsible ML deployment.

To help the reader navigate the remainder of this article, we provide a summary of the paper's structure. Section 2 introduces the ten ethical principles and explains their rationale. We build upon this foundation with additional principles and emerging considerations, including context-sensitive fairness, continuous monitoring, environmental stewardship, privacy-preserving unlearning, and stakeholder engagement. Section 3 examines improper uses and ethical breaches through case studies on mortgage redlining, the Pima Indians' diabetes dataset, and hallucinations in LLMs. Section 4 presents a weighted scoring rubric and algorithmic process for evaluating ML systems, including demonstration cases on mortgage underwriting and diabetes prediction; tables summarize the raw scores, weights, and weighted scores for each example.

Section 5 compares our framework with existing international guidelines, while Section 6 discusses ecological aspects, including energy consumption, water withdrawal, planetary health, and the debate on nuclear dependence. Section 7 outlines governance and accountability considerations; Section 8 proposes a roadmap and recommendations for practitioners; Section 9 suggests avenues for future research; and Section 10 concludes the paper.

## 2 Ethical Principles for Responsible ML

Ethical principles provide a conceptual foundation for evaluating the behavior of ML systems. However, recent scholarship notes that many existing AI ethics guidelines remain high-level and heterogeneous across application domains. A recent scoping review of fairness, accountability, transparency, and ethics (FATE) in AI for social media and healthcare highlights the heterogeneity of guidelines and calls for the development of systematic evaluation frameworks, [2]. Reviews of fairness in healthcare note that algorithmic biases can lead to misdiagnosis, unequal access to treatment, and other ethical harms, [14]. To address these challenges, the following subsections introduce ten principles—accuracy, bias and fairness, accessibility, security, privacy, transparency, accountability, human oversight, sustainability, and harm avoidance—and explain their rationale and interconnections. Each principle is motivated by real-world concerns and forms part of the framework developed in this paper.

### 2.1 Overview of the Ten Principles

1. Accuracy. ML systems should provide consistent, reliable outputs within their intended domain. Accurate models are critical in high-stakes domains such as credit underwriting, [15]; thus, data and models must be validated to ensure that predictions accurately reflect the ground truth without unacceptable error margins.
2. Bias and fairness. Datasets and algorithms should be audited for disparate impact. The OECD emphasizes human-centered values and fairness, urging organizations to prevent discrimination by using diverse training data and conducting regular bias audits, [12]. The Pima Indians dataset illustrates how imbalanced class distributions can lead to biased classifiers. At the same time, recent experiments with chatbots in mortgage decision-making show 8.5 % more approvals for white applicants than for identical Black applicants, [16].
3. Accessibility and inclusivity. Systems should be designed for diverse users, including those with

disabilities or limited technical literacy, to ensure accessibility and inclusivity. UNESCO calls for awareness, literacy, and inclusive participation in AI design, [11]. Datasets must represent the intended population; failure to do so can exacerbate digital divides.

4. Security and resilience. AI systems must be robust against adversarial attacks and resilient to failures. The OECD stresses robustness, security, and safety through stress tests and cybersecurity measures, [12]. Design should include logging and traceability to support post-hoc audits, [9].
5. Privacy and data protection. Protection of personal data is paramount. UNESCO's recommendation highlights the right to privacy and data protection, [11]. Data minimization and techniques such as differential privacy should be employed, and models should be examined for risks associated with training data extraction, [17].
6. Transparency and explainability. Systems should provide understandable explanations of their decisions. The EU AI Act requires transparency obligations for high-risk AI, [9], while the OECD calls for transparent and explainable AI operations, [12]. LLM hallucinations demonstrate the hazards of opaque models, [1].
7. Accountability and governance. Organizations and individuals deploying AI must be accountable for outcomes. The OECD emphasizes legal, ethical, and operational accountability, [12]. NIST also emphasizes the importance of documentation, risk management, and human oversight, [10].
8. Human oversight and autonomy. AI should assist rather than replace human decision-makers. UNESCO urges human determination and oversight, [11]. Systems must include mechanisms for human review and intervention to avoid overreliance on automation.
9. Sustainability and environmental impact. The ecological footprint of AI should be minimized. Training LLMs can emit hundreds of tonnes of CO<sub>2</sub> [18], [19], prompting calls for Green AI research to make efficiency a core evaluation metric, [20]. Data centers could consume up to 9 % of U.S. electricity by 2030, [21], raising questions about the energy sources powering AI. We discuss the "nuclear dependence" argument later.

10. Harm avoidance and safety. Systems must be designed to prevent harm to individuals and society. UNESCO's principle of proportionality and do-no-harm, [11], aligns with this requirement. Redlining, dataset bias, and hallucinations illustrate how unethical ML can cause discrimination and misinformation.

## 2.2 Connections to the Ten Commandments of Computer Ethics

The Ten Commandments of Computer Ethics emphasize the importance of avoiding harm, respecting privacy, and considering the societal consequences, [13]. These enduring principles underlie our framework: for example, the commandment "thou shalt not use a computer to harm other people" aligns with our harm-avoidance and fairness principles; "thou shalt always use a computer in ways that ensure consideration and respect" relates to accessibility and human oversight; and "thou shalt think about the social consequences of the programs you write" echoes accountability and transparency.

There is a simplified causal pathway from misuse to social harm. When practitioners deploy unvetted datasets or ignore model drift, ethical breaches such as bias, privacy violations, or opacity occur. These breaches can lead to discrimination, undermine public trust, and cause safety hazards. Understanding this chain of events highlights the importance of preventive audits and accountability mechanisms, [22].

## 2.3 Additional Principles and Considerations

The ten principles outlined above provide a strong foundation for ethical ML, but recent research highlights additional considerations that deserve explicit attention. This subsection introduces complementary principles that expand the ethical horizon beyond the traditional ten.

**Context-sensitive fairness.** Fairness definitions are not one-size-fits-all. Biases can arise from domain-specific factors, such as genetic variation and data acquisition, in the medical field. They caution that simply omitting protected attributes does not guarantee fairness; underrepresented groups can be underdiagnosed, and unequal access to care may persist even when sensitive variables are removed, [3]. A context-sensitive approach tailors fairness metrics and mitigation techniques to the specific harms and populations affected. For example, loan underwriting should evaluate the disparate impacts across intersecting identities (such as race, gender,

and income), while healthcare applications must account for genetic diversity and socio-economic determinants.

**Continuous monitoring and MLOps.** Fairness and ethical quality can degrade over time as data distributions shift or new user groups emerge. Practitioners lack tools and processes for fairness-aware engineering and call for MLOps practices that continuously audit, monitor, and improve fairness across the ML lifecycle, [4]. Continuous monitoring entails tracking performance metrics disaggregated by subgroup, implementing drift detection, and automating retraining or model updates when ethical thresholds are breached.

**Environmental stewardship and digital planetary health.** AI development has material consequences for ecosystems. Lupton frames generative AI within a digital planetary health perspective, emphasizing that AI infrastructure produces electronic waste, greenhouse gas emissions, and water consumption, and argues for multispecies justice and intergenerational considerations, [6]. Global AI water withdrawal could reach billions of cubic meters by 2027, prompting a call to reduce the technology's hidden water footprint [8]. An expanded sustainability principle, therefore, encompasses carbon, water, rare-earth minerals, and the broader ecological impacts of AI.

**Stakeholder engagement and social justice.** Ethical ML must involve the communities it affects. The digital planetary health perspective emphasizes the recognition of the rights and voices of both human and non-human stakeholders, [6]. Engaging patients, borrowers, community advocates, and domain experts throughout the ML lifecycle fosters trust, surfaces contextual values, and helps align system behavior with societal goals. These considerations extend beyond the conventional ten principles by foregrounding social justice and multispecies ethics.

### 3 Improper Uses and Ethical Breaches: Case Studies

Understanding how ethical principles can be violated requires concrete examples. This section presents case studies drawn from journalism and academic research—such as mortgage redlining, dataset bias, hallucinations in LLMs, and ecological considerations—to illustrate how misuse, design oversights, or opaque models can produce social harm. Systematic reviews of LLMs in surgical research identify accuracy, bias, confidentiality, and responsibility as prevalent ethical concerns,

[23]. Analyses of public health AI suggest that models trained on data from wealthy regions may systematically misdiagnose or underdiagnose underserved populations, thereby exacerbating health disparities, [24]. Biases can be amplified when AI systems are proprietary and closed source, making it challenging to trace unfair decisions. In healthcare, bias can originate from both the data used to train models and from the algorithms themselves. Scholars emphasize that addressing explainability and algorithmic bias is not only a technical challenge but also involves fairness and trust, requiring international rules and regulations to ensure the ethical use of AI, [25]. By examining these scenarios, we highlight the consequences of ignoring ethical principles and underscore the importance of audits and accountability mechanisms.

#### 3.1 Algorithmic Redlining in Mortgage Lending

Investigative journalists and researchers have uncovered discriminatory behavior in mortgage algorithms. A markup investigation found that, after controlling for applicants' financial variables, applicants of color were 40–80 % more likely to be denied loans than white applicants; disparities exceeded 250 % in some metropolitan areas. A Lehigh University experiment fed 6,000 synthetic mortgage applications to commercially available chatbots. The chatbots recommended loan denials more often for Black applicants than for identical white applicants, approved white applicants 8.5 % more often, and offered Black applicants higher interest rates. The study concluded that race proxies, such as credit scores and zip codes, enable algorithms to replicate historical redlining, [16]. These findings underscore the need to audit training data and model outcomes for disparate impact.

#### 3.2 Dataset Bias: The Pima Indians Diabetes Database

The Pima Indians diabetes database contains data on 768 female patients of Pima heritage aged 21 years or older, with eight clinical variables. Only 268 patients are positive for diabetes, while 500 are negative; thus, the dataset is imbalanced. Models trained on such data may optimize for overall accuracy by predicting the majority class, leading to high false-negative rates among diabetic patients. Moreover, the dataset comprises only one ethnic group, raising concerns about the generalizability of models to broader populations. This case illustrates the importance of representative sampling and balanced data in fairness audits.

### 3.3 Hallucinations in Large Language Models

LLMs sometimes generate statements not grounded in data. In clinical contexts, LLM hallucinations can fabricate medical facts, generate false references, or misrepresent patient information. Examples include ChatGPT inventing citations about liver involvement in Pompe disease and the Whisper speech-to-text model hallucinating violent or racially charged remarks, [1]. In 2023, U.S. attorneys were sanctioned \$5,000 for submitting a legal brief containing six fictitious court cases produced by ChatGPT, [26]. Such hallucinations highlight the importance of verification, retrieval-augmented generation, and human oversight during deployment.

A recent multi-model assurance analysis evaluated six LLMs on thousands of clinical cases and found that adversarial prompt injections can trigger hallucination rates between 50 % and 82 %, even when models are prompted to provide references, [27]. Mitigation strategies reduced hallucinations only modestly, underscoring the need for safeguards and verification mechanisms.

### 3.4 Ecological and Nuclear Dependence Considerations

Training modern LLMs can emit hundreds of tonnes of carbon dioxide: the environmental cost of one transformer training run has been estimated at 284 tCO<sub>2</sub>, compared with the 5 t annual carbon footprint of an average person, [18], [19]. The Electric Power Research Institute projects that data centers could consume up to 9 % of U.S. electricity by 2030, [21]. Some commentators argue that to meet this demand, AI will depend on reliable, low-carbon baseload energy such as nuclear power. While atomic energy offers high reliability, it raises concerns regarding safety, waste, and geopolitics. These debates underscore the importance of our sustainability principle and call for research into energy-efficient architectures, such as Green AI, [20], as well as the integration of renewable energy.

Beyond carbon emissions, AI development has hidden environmental costs. A digital planetary health perspective emphasizes that the construction and operation of data centers generate large amounts of electronic waste and consume scarce water resources. Lupton argues that generative AI increases e-waste, greenhouse gas emissions, and water use and calls for multispecies justice and intergenerational responsibility in technology policy, [6]. Li and colleagues estimate that AI's global water withdrawal could reach 4.2–6.6 billion cubic meters by 2027 and quantify the water footprint of training LLMs, [8]. Penn State researchers

predict that data centers could account for up to 20 % of global electricity consumption by 2030–2035, amplifying pressure on electric grids and water resources. These findings support our sustainability principle and highlight the need for energy-efficient algorithms, renewable-powered data centers, and governance that considers both carbon and water footprints. Environmental stewardship should therefore incorporate not only Green AI practices but also water-aware scheduling, hardware recycling, and cross-industry collaboration to reduce AI's hidden footprints.

## 4 Ethical Scoring Rubric and Algorithmic Process

High-level ethical principles alone are insufficient without a means of operationalizing them. Recent literature emphasizes the heterogeneity of existing guidelines and argues for the development of systematic tools that measure fairness, accountability, transparency, and ethics. This section introduces a weighted scoring rubric and an algorithmic process for assessing ML systems. The rubric quantifies compliance with each principle using a set of weights and thresholds. At the same time, the process describes how to audit data and models, assign scores, aggregate results, and iterate on improvements.

### 4.1 Designing a Weighted Rubric

To operationalize the ethical principles, we propose a rubric that assigns weights and scores to each principle, thereby providing a framework for evaluating the moral implications of decisions. The rubric enables practitioners to diagnose deficiencies and prioritize improvements. Table 1 presents an example weighting scheme. We assign higher weights to principles that directly affect human rights (bias/fairness, harm avoidance, and privacy) and slightly lower weights to auxiliary properties (accessibility and sustainability). Each principle is rated on a 0–5 scale (0 = unacceptable, 5 = excellent). Thresholds for compliance are defined as follows: High (weighted score  $\geq 80$ ) indicates the system meets best practices, Medium (60–79) denotes satisfactory compliance with minor issues, and Low (<60) signals unacceptable ethical risk.

Table 1: Illustrative ethical scoring rubric with weights and descriptors. Higher weights reflect direct effects on human rights (bias/fairness, harm avoidance, privacy), while lower weights apply to auxiliary properties (accessibility, sustainability). Source: created by the author.

Principle	Description	Weight
Accuracy	Validity and reliability of model outputs, [15]	10
Bias/Fairness	Mitigation of disparate impact; fairness audits [12]	12
Accessibility	Inclusive design for diverse user groups, [11]	6
Security	Resilience to adversarial manipulation and misuse, [12]	8
Privacy	Protection, minimization, and secure handling of data, [11]	10
Transparency	Explainability, traceability, documentation, [12]	8
Accountability	Governance, audit trails, redress mechanisms, [12]	8
Human oversight	Human-in-the-loop safeguards and escalation paths [11]	6
Sustainability	Energy efficiency and environmental impact, [18]	6
Harm avoidance	Prevention of individual and societal harm, [11]	10

## 4.2 Algorithmic Process for Ethical Assessment

To apply the rubric systematically, we propose the following algorithmic process:

1. Context definition. Define the ML system’s purpose, domain, and stakeholders. Identify the potential harms and benefits.
2. Data and model audit. Examine datasets for representativeness and imbalance, [28]. Evaluate models for error rates across subgroups. Document data provenance using datasheets, [29] and model cards, [30].

3. Score assignment. For each principle, assign a raw score (0–5) based on evidence from audits, documentation, and stakeholder feedback. Multiply the raw score by the weight from Table 1.
4. Aggregate and classify. Sum the weighted scores to obtain the overall ethical score. Classify compliance as High ( $\geq 80$ ), Medium (60–79), or Low ( $<60$ ).
5. Mitigation and iteration. For principles with low scores, propose mitigation strategies such as rebalancing data, integrating human oversight, or adopting energy-efficient architectures, [28], [20]. Repeat the audit after mitigation to track improvements.

## 4.3 Demonstration Cases: Mortgage Underwriting and Diabetes Prediction

Consider a simplified mortgage-underwriting model trained on historical loan applications. The audit reveals that the dataset contains proxies for race (such as zip codes and credit scores) and is biased towards past applicants; redlining patterns emerge in the model’s predictions, [16]. Suppose we assign raw scores as follows: Accuracy = 4 (validated on hold-out data), Bias/Fairness = 1 (significant disparate impact), Accessibility = 3 (user interface supports multiple languages but lacks accommodations for disabled users), Security = 4 (models are robust to adversarial examples), Privacy = 3 (basic data anonymization but no differential privacy), Transparency = 2 (limited explainability), Accountability = 3 (basic logging but unclear governance roles), Human oversight = 3 (loan officers can override automated decisions), Sustainability = 2 (energy consumption not measured), and Harm avoidance = 2 (evidence of discrimination). Using the weights in Table 1, the overall weighted score is  $4 \times 10 + 1 \times 12 + 3 \times 6 + 4 \times 8 + 3 \times 10 + 2 \times 8 + 3 \times 8 + 3 \times 6 + 2 \times 6 + 2 \times 10 = 222$  out of a maximum of 500, yielding a compliance percentage of 44.4 % (Low). As seen in Table 2, this diagnostic indicates the need for urgent action to address bias, transparency, harm avoidance, and sustainability. After applying fairness mitigation techniques (such as reweighing or adversarial debiasing), instituting model cards and datasheets, and measuring energy consumption, the model can be reassessed, and its score can improve, [31].

Table 2: Breakdown of the mortgage underwriting ethical score. Raw scores from the demonstration are combined with weights from Table 1. Source: created by the author.

Principle	Weight	Raw score	Weighted score
Accuracy	10	4	40
Bias/Fairness	12	1	12
Accessibility	6	3	18
Security	8	4	32
Privacy	10	3	30
Transparency	8	2	16
Accountability	8	3	24
Human oversight	6	3	18
Sustainability	6	2	12
Harm avoidance	10	2	20
<b>Total</b>			<b>222</b>

**Diabetes prediction example** To illustrate the rubric on a real machine-learning model, we trained a logistic regression classifier on the Pima Indians diabetes dataset. The data contain eight clinical attributes from 768 patients, but only 268 positive cases, resulting in a highly imbalanced dataset. The classifier achieved an overall accuracy of about 75 % on a hold-out set but exhibited high false-negative rates for diabetic patients. We assigned raw scores based on audit findings: Accuracy = 3 (moderate prediction quality), Bias/Fairness = 2 (evidence of minority bias and imbalanced data), Accessibility = 4 (simple interface), Security = 3, Privacy = 3, Transparency = 4 (model coefficients are interpretable), Accountability = 3 (basic documentation), Human oversight = 3 (clinician review available), Sustainability = 3 (small model and limited energy usage), and Harm avoidance = 2 (risk of misdiagnosis). Using the weights in Table 1, the weighted score totals roughly 300 out of 500 (60 %), classified as Medium. As seen in Table 3, this diagnostic highlighted fairness and harm avoidance as key areas for improvement. After oversampling the minority class and retraining, the fairness score improved to 4, and the overall compliance rose to about 72 %. This practical example demonstrates how the rubric guides the identification of ethical deficiencies and informs mitigation strategies.

Table 3: Breakdown of the diabetes prediction ethical score. Raw scores reflect the initial assessment; weights are taken from Table 1. Source: created by the author.

Principle	Weight	Raw score	Weighted score
Accuracy	10	3	30
Bias/Fairness	12	4	48
Accessibility	6	4	24
Security	8	4	32
Privacy	10	4	40
Transparency	8	4	32
Accountability	8	3	24
Human oversight	6	3	18
Sustainability	6	4	24
Harm avoidance	10	3	30
<b>Total</b>			<b>302</b>

## 5 Comparison with Existing Ethical Frameworks

No single ethical guideline covers all dimensions of responsible AI. The following comparison examines how our ten principles align with the EU AI Act, NIST’s AI Risk Management Framework, UNESCO’s recommendations, and the OECD AI Principles, highlighting areas of overlap and omission. Scholars argue that relying on voluntary codes or company-led ethics programs is insufficient; trustworthiness requires transparency, robust regulation, and international cooperation. Ethics scholars emphasize that explainability and bias cannot be delegated solely to developers and that international rules and regulations are needed to ensure fairness and trust, [25]. Clinicians also report that accountability, transparency, and bias remain unresolved in AI-assisted decision-making and call for inclusive datasets and clear regulatory frameworks, [32]. Narrative reviews of AI governance further highlight transparency and accountability as essential principles and propose technical, legal, ethical, and interdisciplinary approaches to realize them, [33].

Table 4 compares our framework against major AI governance guidelines. The EU AI Act employs a risk-based classification, imposing stringent obligations on high-risk systems, including the use of high-quality datasets and robust risk management. NIST’s AI RMF emphasizes the characteristics of trustworthy AI, including validity, reliability, safety, security, accountability, transparency, explainability, privacy, and fairness. UNESCO’s

Table 4: Comparison of the proposed framework with selected AI ethics guidelines. ✓ indicates explicit coverage, — denotes implicit coverage or omission. Source: created by the author.

Principle	EU AI Act	NIST	UNESCO	OECD	Proposed
Accuracy	✓	✓	✓	✓	✓
Bias/Fairness	✓	✓	✓	✓	✓
Accessibility	—	—	✓	✓	✓
Security	✓	✓	✓	✓	✓
Privacy	✓	✓	✓	✓	✓
Transparency	✓	✓	✓	✓	✓
Accountability	✓	✓	✓	✓	✓
Human oversight	✓	✓	✓	—	✓
Sustainability	—	—	✓	✓	✓
Harm avoidance	—	—	✓	—	✓
Scoring rubric	—	—	—	—	✓

recommendations list ten principles, including proportionality, safety, privacy, multi-stakeholder governance, responsibility, transparency, human oversight, sustainability, awareness, and fairness. The OECD AI Principles call for inclusive growth, human-centered values, transparency, robustness, and accountability. Our framework incorporates these elements, but it also adds explicit scoring, weighting, and a structured assessment process. It also integrates sustainability and harm avoidance as distinct principles, emphasizing the Ten Commandments of Computer Ethics. This combination offers a holistic and actionable approach that is not found in any single existing framework.

## 6 Ecological Aspects and the Nuclear Dependence Argument

Ethical AI must also consider environmental sustainability. Data centers and ML training consume significant energy, contributing to greenhouse gas emissions, and there is debate about whether AI development will increase dependence on nuclear power. Transparency and accountability must extend beyond algorithmic decisions to encompass the provenance of energy sources, the carbon footprint of models, and the broader societal impacts of resource consumption. The following subsections examine the energy and carbon footprint of ML and discuss the arguments for and against incorporating nuclear power into AI infrastructure.

### 6.1 Energy and Carbon Footprint

The environmental impact of ML stems from both training and deployment. Training a transformer model for machine translation produced 284 tCO<sub>2</sub>,

while the average human emits about 5 t per year, [18], [19]. The Green AI movement advocates making efficiency a core evaluation metric and reporting the “price tag” of training and running models, [20]. Our framework encourages researchers to monitor and minimize energy use, adopt model compression and quantization techniques, and explore low-resource algorithms.

### 6.2 Powering AI: Renewables and Nuclear

Meeting the growing electricity demand of data centers, which could reach 9 % of U.S. supply by 2030, will require large-scale, low-carbon power. Renewable sources, such as solar and wind, are intermittent; nuclear energy provides a reliable baseload but introduces safety, waste management, and geopolitical risks, [21]. The “nuclear dependence” argument, therefore, posits that AI’s energy appetite might indirectly promote nuclear power development. However, advances in energy-efficient models, hardware accelerators, and demand-response strategies may reduce this dependence. Policymakers should strike a balance between investments in clean energy infrastructure and oversight to prevent new forms of environmental or social harm.

## 7 Governance, Oversight and Accountability

Establishing clear governance structures is crucial for effectively implementing ethical principles into practice. Clinicians and other stakeholders have expressed concern that opaque, black-box AI systems undermine accountability and perpetuate bias. Researchers caution that trust cannot be grounded solely in industry self-regulation and emphasize the need for international rules and regulatory frameworks to enforce explainability and fairness, [34]. This section discusses governance mechanisms, oversight procedures, and accountability frameworks that support the responsible deployment of ML.

Institutions deploying ML must establish clear governance structures. The EU AI Act requires providers of high-risk AI systems to implement risk management, data governance, technical documentation, and human oversight, [9]. NIST’s AI RMF emphasizes accountability, transparency, and stakeholder engagement, [10]. UNESCO recommends multi-stakeholder governance and adaptive collaboration, [11]. Our framework aligns with these requirements by requiring documentation (datasheets and model cards), [29], [30], audit trails, grievance mechanisms, and the assignment of responsible officers. When AI errors occur, as in the ChatGPT-generated fictitious court cases,

[26], accountability mechanisms ensure that human agents—not the AI—bear legal responsibility.

Empirical studies on fairness-aware engineering highlight additional governance challenges. Fairness is often treated as a second-class quality in software projects, and practitioners lack standardized processes, tooling, and expertise to build fairness into the ML pipeline, [4], [35]. They call for fairness-aware MLOps practices—such as continuous auditing, documentation, and stakeholder engagement—to ensure fairness does not erode as models evolve and improve. Governance structures should therefore include dedicated fairness officers, periodic audits, and transparent reporting on efforts to mitigate bias.

Governance must also anticipate privacy threats. Membership inference attacks enable adversaries to determine whether specific data points were utilized in training a model, thereby exposing individuals to privacy risks. Incorporating privacy-preserving training methods, unlearning protocols, and differential privacy into governance frameworks will help protect data subjects and maintain public trust. By integrating fairness-aware engineering and privacy-by-design into oversight processes, organizations can operationalize ethical principles and strengthen accountability.

## 8 Roadmap and Recommendations

Building on the preceding analysis, this section outlines a roadmap of practical steps for organizations seeking to operationalize responsible ML. The recommendations emphasize the importance of structured documentation, fairness auditing, transparency, human oversight, sustainability, accountability, and stakeholder engagement. These actions are synthesized from our framework and the literature and are intended to guide developers, deployers, and regulators toward the ethical implementation of these technologies.

The evolution of AI ethics frameworks has accelerated in recent years, with milestones including the IEEE 7000 standard (2017), the EU AI Act (2024), and the frameworks proposed in this paper. This progression highlights the necessity for ongoing adaptation as new technologies and societal expectations emerge.

Based on our analysis, we recommend the following roadmap:

1. Adopt structured documentation. Use datasheets for datasets [29] and model cards [30] to record the provenance, composition, and performance of data and models.
2. Integrate fairness auditing. Employ metrics such as disparate impact and equalized odds, and

mitigate bias through reweighing, adversarial debiasing, or custom loss functions. Use tools like Fairness Indicators and report results by demographic subgroup.

3. Prioritize transparency and explainability. Incorporate interpretable models when feasible; for opaque models, provide post-hoc explanations and confidence measures. Offer clear user documentation to demystify system behavior, [36].
4. Ensure human oversight. Maintain the human-in-the-loop for high-stakes decisions; implement override mechanisms and require second opinions for automated recommendations.
5. Address sustainability. Evaluate the energy consumption of models and infrastructure; favor Green AI practices, [20], and consider carbon offsetting. Encourage investment in renewable and safe energy sources while debating the role of nuclear energy.
6. Establish accountability. Define roles and responsibilities for developers, deployers, and regulators. Implement grievance procedures and auditing committees to monitor compliance and handle incidents.
7. Educate and involve stakeholders. Provide ethics and bias training for developers and end users; engage diverse stakeholders (patients, borrowers, and community advocates) throughout the ML lifecycle. Education in ethics and core AI concepts for developers, leaders, and users is critical for responsible AI adoption, [37]. Practitioners should also address the risks of misinformation, disinformation, and hallucinations highlighted in the broader human–AI discourse, [38], and develop resources to reduce anxiety and promote responsible use.
8. Implement continuous monitoring and MLOps. Adopt fairness-aware engineering practices that include ongoing auditing of model performance across demographic groups, drift detection, and automated retraining when ethical thresholds are breached. Fairness is often treated as a secondary concern. MLOps processes should embed fairness into the ML lifecycle, [4].
9. Adopt privacy-preserving training and unlearning. Mitigate membership inference attacks by incorporating differential privacy, secure federated learning, and dual-purpose

training methods such as DuoLearn, [5]. Establish procedures for unlearning when individuals withdraw consent, and ensure that models can be updated without exposing training data.

10. Embrace planetary health and environmental stewardship. Recognize that AI development impacts not only carbon emissions but also water use, electronic waste, and biodiversity. Employ digital planetary health principles to evaluate the full ecological footprint of models and infrastructure, and invest in renewable energy, water-aware scheduling, and hardware recycling, [6]. These actions complement the sustainability principle and align AI progress with environmental justice.

## 9 Future Research Directions

Research into responsible ML is evolving quickly. Several avenues warrant further investigation. First, the development of context-sensitive fairness metrics and mitigation techniques remains an open problem. Existing metrics, such as disparate impact and equalized odds, treat protected attributes uniformly across domains; however, practical applications in finance, healthcare, and education require nuanced definitions of harm and fairness. Future work should explore adaptive metrics that account for intersectional identities and dynamic feedback effects, as well as techniques to mitigate bias without sacrificing accuracy.

Second, the energy footprint of AI systems calls for novel algorithmic and architectural solutions. Green AI initiatives emphasize efficiency, yet the training and deployment of LLMs continue to consume significant resources. Researchers should focus on model compression, sparsity, transfer learning, and hardware-aware optimization. Investigating the environmental trade-offs of different power sources, including renewable and nuclear energy, will also inform sustainable infrastructure planning.

Third, improved frameworks for explainability and trust are crucial. As hallucinations in LLMs illustrate, black-box behaviors can mislead users and erode confidence. Future research should integrate retrieval-augmented generation, verifiable reasoning, and interactive explanation interfaces that adapt to user expertise. Empirical studies on how explanations influence human decision-making will help calibrate transparency mechanisms, [39].

Finally, responsible ML governance requires continuous adaptation. Comparative analyses of emerging regulations across jurisdictions (EU, U.S., UNESCO, OECD, and others) will provide insights

into harmonization and gaps. The interplay between ethical principles and legal requirements should be examined to guide policymakers and practitioners. Longitudinal studies measuring the effectiveness of scoring rubrics and auditing frameworks in real-world deployments will help refine our proposed approach.

Additional research directions stem from the emerging considerations outlined earlier. Fairness-aware MLOps and continuous monitoring remain under-explored. Future work should develop tools for tracking fairness metrics over time, automatic debiasing routines, and adaptive weighting schemes that maintain equity across shifting populations, [4]. Another frontier is privacy-preserving training and unlearning. Dynamic token selection and unlearning can mitigate membership inference while preserving model utility, [5]. Extending these ideas to multimodal models, hybrid architectures, and large-scale deployment poses an open challenge.

Environmental sustainability also warrants further investigation. The digital planetary health perspective suggests broadening sustainability metrics beyond carbon emissions to include water consumption, electronic waste, and biodiversity impacts. The hidden water footprint of AI training forecasts massive water withdrawals in the coming years, [8]. Future work should develop water-aware scheduling algorithms, energy–water trade-off models, and circular economy approaches for AI hardware. Cross-cultural and multispecies justice perspectives could inform ethical frameworks that address AI’s ecological and social footprint across global contexts.

## 10 Conclusion

Responsible ML requires more than aspirational statements; it demands concrete frameworks, tools, and cultural change. By synthesizing ten ethical principles, proposing a weighted scoring rubric, examining case studies of redlining, dataset bias, and hallucinations, and integrating sustainability and governance considerations, this paper provides a comprehensive roadmap for ethical ML deployment. Our comparison with existing frameworks shows complementarity while highlighting the novelty of our scoring approach. The coming years will see the continued evolution of AI governance. Adopting structured documentation, fairness auditing, transparency, human oversight, and sustainable practices will ensure that ML benefits society without exacerbating harm.

Our demonstration cases illustrate how the weighted rubric can identify deficiencies and inform improvements. The mortgage example revealed that bias and harm avoidance remain critical issues,

while the diabetes case underscored the importance of fairness, transparency, and interpretability. The accompanying tables make the mathematics transparent by breaking down raw scores, weights, and totals, enabling practitioners to audit their systems systematically. These examples illustrate that ethical compliance is not a binary state but a spectrum that can be quantified, diagnosed, and iteratively improved.

Looking ahead, integrating fairness-aware MLOps, privacy-preserving training methods such as DuoLearn, and planetary health metrics promises to advance responsible AI practices. By recognizing the importance of continuous monitoring, privacy by design, environmental stewardship, and stakeholder engagement, practitioners and policymakers can build AI systems that safeguard human rights, protect sensitive data, and sustain the planet. We hope this work inspires researchers, developers, and regulators to embrace a broadened ethical agenda that addresses emerging challenges and opportunities in the rapidly evolving field of ML.

#### References:

- [1] M. Ashwin, S. Jha, G. Prasad, and S. Kumar, "Fake it till you make it? ai hallucinations and ethical dilemmas in anesthesia research and practice," *Journal of Anaesthesiology Clinical Pharmacology*, vol. 41, no. 3, p. 381–383, Jun. 2025. [Online]. Available: [http://dx.doi.org/10.4103/joacp.joacp\\_56\\_25](http://dx.doi.org/10.4103/joacp.joacp_56_25)
- [2] A. Singhal, N. Neveditsin, H. Tanveer, and V. Mago, "Toward fairness, accountability, transparency, and ethics in ai for social media and health care: Scoping review," *JMIR Medical Informatics*, vol. 12, p. e50048, Apr. 2024. [Online]. Available: <http://dx.doi.org/10.2196/50048>
- [3] J. W. Anderson and S. Visweswaran, "Algorithmic individual fairness and healthcare: a scoping review," *JAMIA Open*, vol. 8, no. 1, Dec. 2024. [Online]. Available: <http://dx.doi.org/10.1093/jamiaopen/ooae149>
- [4] C. Ferrara, G. Sellitto, F. Ferrucci, F. Palomba, and A. De Lucia, "Fairness-aware machine learning engineering: how far are we?" *Empirical Software Engineering*, vol. 29, no. 1, Nov. 2023. [Online]. Available: <http://dx.doi.org/10.1007/s10664-023-10402-y>
- [5] Z. Tong, F. Sun, and L. M. Nguyen, *Pretraining Data Exposure in Large Language Models: A Survey of Membership Inference, Data Contamination, and Security Implications*. Springer Nature Switzerland, Jul. 2025, p. 152–162. [Online]. Available: [http://dx.doi.org/10.1007/978-3-031-97144-0\\_14](http://dx.doi.org/10.1007/978-3-031-97144-0_14)
- [6] D. Lupton, "Towards a digital planetary health perspective: generative ai and the digital determinants of health," *Health Promotion International*, vol. 40, no. 5, Sep. 2025. [Online]. Available: <http://dx.doi.org/10.1093/heapro/daaf153>
- [7] M. Leon, "Generative artificial intelligence and prompt engineering: A comprehensive guide to models, methods, and best practices," *Advances in Science, Technology and Engineering Systems Journal*, vol. 10, no. 02, p. 01–11, Mar. 2025. [Online]. Available: <http://dx.doi.org/10.25046/aj100201>
- [8] P. Li, J. Yang, M. A. Islam, and S. Ren, "Making ai less "thirsty"," *Communications of the ACM*, vol. 68, no. 7, p. 54–61, Jun. 2025. [Online]. Available: <http://dx.doi.org/10.1145/3724499>
- [9] N. T. Nikolinakos, "Eu policy and legal framework for artificial intelligence, robotics and related technologies - the ai act," *Law, Governance and Technology Series*, 2023. [Online]. Available: <http://dx.doi.org/10.1007/978-3-031-27953-9>
- [10] E. Tabassi, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, Jan. 2023. [Online]. Available: <http://dx.doi.org/10.6028/NIST.AI.100-1>
- [11] D. E. van Norren, "The ethics of artificial intelligence, unesco and the african ubuntu perspective," *Journal of Information, Communication and Ethics in Society*, vol. 21, no. 1, p. 112–128, Dec. 2022. [Online]. Available: <http://dx.doi.org/10.1108/JICES-04-2022-0037>
- [12] A. Wodi, "Artificial intelligence (ai) governance: An overview," *SSRN Electronic Journal*, 2024. [Online]. Available: <http://dx.doi.org/10.2139/ssrn.4840769>
- [13] J. Zhou, H. Müller, A. Holzinger, and F. Chen, "Ethical chatgpt: Concerns, challenges, and commandments," *Electronics*, vol. 13, no. 17, p. 3417, Aug. 2024. [Online]. Available: <http://dx.doi.org/10.3390/electronics13173417>
- [14] D. Ueda, T. Kakinuma, S. Fujita, K. Kamagata, Y. Fushimi, R. Ito, Y. Matsui, T. Nozaki, T. Nakaura, N. Fujima, F. Tatsugami, M. Yanagawa, K. Hirata, A. Yamada,

- T. Tsuboyama, M. Kawamura, T. Fujioka, and S. Naganawa, "Fairness of artificial intelligence in healthcare: review and recommendations," *Japanese Journal of Radiology*, vol. 42, no. 1, p. 3–15, Aug. 2023. [Online]. Available: <http://dx.doi.org/10.1007/s11604-023-01474-3>
- [15] G. Pennisi, "Operationalization of (trans)gender in facial recognition systems: From binarism to intersectionality," *Future Humanities*, vol. 2, no. 3, Jul. 2024. [Online]. Available: <http://dx.doi.org/10.1002/fhu2.17>
- [16] A. Mergen, N. undefinedetin Kılıç, and M. F. Özbilgin, *Artificial Intelligence and Bias Towards Marginalised Groups: Theoretical Roots and Challenges*. Emerald Publishing Limited, Apr. 2025, p. 17–38. [Online]. Available: <http://dx.doi.org/10.1108/S2051-233320250000012004>
- [17] J. Dagdelen, A. Dunn, S. Lee, N. Walker, A. S. Rosen, G. Ceder, K. A. Persson, and A. Jain, "Structured information extraction from scientific text with large language models," *Nature Communications*, vol. 15, no. 1, Feb. 2024. [Online]. Available: <http://dx.doi.org/10.1038/s41467-024-45563-x>
- [18] M. Han, I. Canli, J. Shah, X. Zhang, I. G. Dino, and S. Kalkan, "Perspectives of machine learning and natural language processing on characterizing positive energy districts," *Buildings*, vol. 14, no. 2, p. 371, Jan. 2024. [Online]. Available: <http://dx.doi.org/10.3390/buildings14020371>
- [19] N. J. Abernethy, "Let stochastic parrots squawk: why academic journals should allow large language models to coauthor articles," *AI and Ethics*, vol. 5, no. 5, p. 4535–4553, Sep. 2024. [Online]. Available: <http://dx.doi.org/10.1007/s43681-024-00575-7>
- [20] V. Bolón-Canedo, L. Morán-Fernández, B. Cancela, and A. Alonso-Betanzos, "A review of green artificial intelligence: Towards a more sustainable future," *Neurocomputing*, vol. 599, p. 128096, Sep. 2024. [Online]. Available: <http://dx.doi.org/10.1016/j.neucom.2024.128096>
- [21] Y. Chen, R. Zhang, J. Lyu, and Y. Hou, "Ai and nuclear: A perfect intersection of danger and potential?" *Energy Economics*, vol. 133, p. 107506, May 2024. [Online]. Available: <http://dx.doi.org/10.1016/j.eneco.2024.107506>
- [22] M. Leon, "Ai safety practices and public perception: Historical analysis, survey insights, and a weighted scoring framework," *Intelligent Systems with Applications*, vol. 28, p. 200583, Dec. 2025. [Online]. Available: <http://dx.doi.org/10.1016/j.iswa.2025.200583>
- [23] S. M. Pressman, S. Borna, C. A. Gomez-Cabello, S. A. Haider, C. Haider, and A. J. Forte, "Ai and ethics: A systematic review of the ethical considerations of large language model use in surgery research," *Healthcare*, vol. 12, no. 8, p. 825, Apr. 2024. [Online]. Available: <http://dx.doi.org/10.3390/healthcare12080825>
- [24] S. A. Haider, S. Borna, C. A. Gomez-Cabello, S. M. Pressman, C. R. Haider, and A. J. Forte, "The algorithmic divide: A systematic review on ai-driven racial disparities in healthcare," *Journal of Racial and Ethnic Health Disparities*, Dec. 2024. [Online]. Available: <http://dx.doi.org/10.1007/s40615-024-02237-0>
- [25] S. Nasir, R. A. Khan, and S. Bai, "Ethical framework for harnessing the power of ai in healthcare and beyond," *IEEE Access*, vol. 12, p. 31014–31035, 2024. [Online]. Available: <http://dx.doi.org/10.1109/ACCESS.2024.3369912>
- [26] A. Joseph, P. Abril, and A. Del Riego, "Chatgpt, esq.: Recasting unauthorized practice of law in the era of generative ai," *SSRN Electronic Journal*, 2025. [Online]. Available: <http://dx.doi.org/10.2139/ssrn.5152523>
- [27] M. Omar, V. Sorin, J. D. Collins, D. Reich, R. Freeman, N. Gavin, A. Charney, L. Stump, N. L. Bragazzi, G. N. Nadkarni, and E. Klang, "Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support," *Communications Medicine*, vol. 5, no. 1, Aug. 2025. [Online]. Available: <http://dx.doi.org/10.1038/s43856-025-01021-3>
- [28] T. D. Jui and P. Rivas, "Fairness issues, current approaches, and challenges in machine learning models," *International Journal of Machine Learning and Cybernetics*, vol. 15, no. 8, p. 3095–3125, Jan. 2024. [Online]. Available: <http://dx.doi.org/10.1007/s13042-023-02083-2>
- [29] C. J. Connolly, D. M. Hueholt, and M. A. Burt, "Datasheets for earth science datasets," *Bulletin of the American Meteorological Society*, vol. 106, no. 4, p. E642–E648, Apr. 2025. [Online]. Available: <http://dx.doi.org/10.1175/BAMS-D-24-0203.1>

- [30] I. Hupont, D. Fernández-Llorca, S. Baldassarri, and E. Gómez, “Use case cards: a use case reporting framework inspired by the european ai act,” *Ethics and Information Technology*, vol. 26, no. 2, Mar. 2024. [Online]. Available: <http://dx.doi.org/10.1007/s10676-024-09757-7>
- [31] M. Leon, “The escalating ai’s energy demands and the imperative need for sustainable solutions,” *WSEAS TRANSACTIONS ON SYSTEMS*, vol. 23, p. 444–457, Dec. 2024. [Online]. Available: <http://dx.doi.org/10.37394/23202.2024.23.46>
- [32] C. Y. Elgin and C. Elgin, “Ethical implications of ai-driven clinical decision support systems on healthcare resource allocation: a qualitative study of healthcare professionals’ perspectives,” *BMC Medical Ethics*, vol. 25, no. 1, Dec. 2024. [Online]. Available: <http://dx.doi.org/10.1186/s12910-024-01151-8>
- [33] B. C. Cheong, “Transparency and accountability in ai systems: safeguarding wellbeing in the age of algorithmic decision-making,” *Frontiers in Human Dynamics*, vol. 6, Jul. 2024. [Online]. Available: <http://dx.doi.org/10.3389/fhumd.2024.1421273>
- [34] H. DeSimone, “Explainable ai: The quest for transparency in business and beyond,” in *2024 7th International Conference on Information and Computer Technologies (ICICT)*. IEEE, Mar. 2024, p. 532–538. [Online]. Available: <http://dx.doi.org/10.1109/ICICT62343.2024.00093>
- [35] M. Leon, G. Napoles, M. M. García, R. Bello, and K. Vanhoof, *Two Steps Individuals Travel Behavior Modeling through Fuzzy Cognitive Maps Pre-definition and Learning*. Springer Berlin Heidelberg, 2011, p. 82–94. [Online]. Available: [http://dx.doi.org/10.1007/978-3-642-25330-0\\_8](http://dx.doi.org/10.1007/978-3-642-25330-0_8)
- [36] G. Napoles, “Prolog-based agnostic explanation module for structured pattern classification,” *Information Sciences*, vol. 622, p. 1196–1227, Apr. 2023. [Online]. Available: <http://dx.doi.org/10.1016/j.ins.2022.12.012>
- [37] G. Biagini, “Towards an ai-literate future: A systematic literature review exploring education, ethics, and applications,” *International Journal of Artificial Intelligence in Education*, vol. 35, no. 4, p. 2616–2666, Mar. 2025. [Online]. Available: <http://dx.doi.org/10.1007/s40593-025-00466-w>
- [38] P. Spitzer, J. Holstein, K. Morrison, K. Holstein, G. Satzger, and N. Kühn, “Don’t be fooled: The misinformation effect of explanations in human–ai collaboration,” *International Journal of Human–Computer Interaction*, p. 1–29, Nov. 2025. [Online]. Available: <http://dx.doi.org/10.1080/10447318.2025.2574511>
- [39] M. Leon, “Gpt-5 and open-weight large language models: Advances in reasoning, transparency, and control,” *Information Systems*, vol. 136, p. 102620, Feb. 2026. [Online]. Available: <http://dx.doi.org/10.1016/j.is.2025.102620>

#### **Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)**

The author contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

#### **Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself**

No funding was received for conducting this study.

#### **Conflict of Interest**

The author has no conflict of interest to declare that is relevant to the content of this article.

#### **Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)**

This article is published under the terms of the Creative Commons Attribution License 4.0

[https://creativecommons.org/licenses/by/4.0/deed.en\\_US](https://creativecommons.org/licenses/by/4.0/deed.en_US)