International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 25s,2025
https://theaspd.com/index.php

Systemic Ethical Dilemmas In Machine Learning: From Predictive Accuracy To Collective Fairness

Martina Gaisch[1]
[1]School of Informatics, Communications & Media, FH Upper Austria, Hagenberg, Austria.
Email: martina.gaisch@fh-hagenberg.at

Abstract: *Machine learning (ML) systems generate significant societal benefits but also pose ethical challenges when fundamental values conflict. This paper investigates ethical tensions in ML processes, focusing on dilemmas of Artificial Intelligence (AI) such as accuracy versus fairness, privacy versus transparency, and personalization versus solidarity. Drawing on international guidelines, including the EU Ethics Guidelines for Trustworthy AI and the EU AI Act, we analyze how these conflicts manifest across domains such as healthcare, finance, and generative AI. A qualitative study among 18 ML experts highlights practitioners' views on fairness, bias, and accountability, revealing that ethical concerns are perceived as systemic rather than isolated issues. Findings suggest that some dilemmas represent unavoidable trade-offs, while others may be mitigated through innovation or governance. We argue that embedding ethical reflection into ML development, supported by regulatory frameworks and participatory deliberation, is essential for ensuring trustworthy AI. By combining conceptual analysis with empirical evidence, the paper contributes to ongoing debates in computational intelligence, emphasizing the importance of aligning ML systems with human values and societal goals.*
Keywords: *Algorithmic Fairness, AI Ethics, Computational Intelligence, Data Protection, Explainability, Transparency.*

1. INTRODUCTION

This work is particularly relevant to multidisciplinary fields such as environmental sciences, where the deployment of machine learning systems intersects with ecological sustainability, climate policy, and governance. As AI becomes increasingly integrated into environmental monitoring, smart agriculture, energy optimization, and risk modeling, understanding the ethical trade-offs in these applications is essential for promoting equitable and sustainable outcomes. Advances in ML and computational intelligence have transformed industries from healthcare to finance. Alongside these benefits, ML systems also create ethical tensions—conflicts between values, norms, and principles that cannot be simultaneously satisfied. These tensions affect trust, fairness, and accountability, raising questions central to computational intelligence research. The objective of this paper is to investigate ethical tensions in ML, classify them into categories of dilemmas, and assess both technical and governance-based strategies for mitigation.

2. RELATED WORK

AI ethics research has highlighted fairness, accountability, transparency, and data protection as recurring themes in ML [18]. [3] mapped over 80 global guidelines, showing convergence on principles such as fairness and privacy [3]. [2] proposed a unified framework for AI ethics, while [6] analyzed sources of harm across the ML lifecycle. Explainability and fairness remain key debates, with scholars warning against black-box models in critical domains ([5]; [4]). [7]; [18]. further document gender bias in generative AI models, emphasizing the persistence of structural inequalities.

More recent work stresses **contextual fairness**—highlighting that general fairness metrics often miss the nuanced ethical requirements of marginalized communities [11]; [19]. Scholars such as Binns [11] and Dastin et al. [20] critique the abstraction in many fairness models, advocating for contextualized approaches. Moreover, explainability tools such as SHAP and LIME have faced scrutiny for producing oversimplified post-hoc interpretations that may mislead stakeholders [12]; [21].

Auditing practices are gaining traction but remain contested. Narayanan [14] argues that current auditing schemes are insufficient and "symbolic," echoing our interviewee concerns. Floridi et al. [22] suggest combining **soft ethics** (self-regulation) and **hard ethics** (enforceable rules) to balance flexibility with accountability. Recent comparative analyses emphasize that algorithmic governance must be adapted to cultural and legal contexts [13], including perspectives from the Global South [23].; [18].

3. METHODOLOGY

This study adopts a **triangulated mixed-methods** approach that integrates conceptual analysis, comparative policy review, and qualitative expert interviews to explore systemic ethical tensions in ML. The conceptual component draws on the **Z-inspection® methodology,** which enables value-sensitive assessment of AI systems through contextual analysis, ethical reflection, and stakeholder deliberation. This framework provided a structured lens for identifying recurring dilemmas—such as fairness versus accuracy or privacy versus transparency—across domains including healthcare, finance, and generative AI. These tensions were not analyzed in isolation but as embedded within the broader sociotechnical fabric of ML deployment.

Complementing the conceptual inquiry, we conducted a **normative review of international AI governance frameworks**, including the **EU Ethics Guidelines for Trustworthy AI** [1], the **AI Act** [8], the **UNESCO Recommendation on the Ethics of AI** [7], and emerging strategies from the **African Union, Brazil, and India** [23]. These documents served as **ethical baselines** to benchmark practitioner perceptions, particularly around core principles such as human agency, accountability, sustainability, and inclusivity. We focused on how regional legal traditions and cultural contexts shape governance strategies, thereby informing our classification of ethical dilemmas not just as technical challenges but as policy-relevant value conflicts.

To ground our analysis in empirical insights, we conducted **semi-structured interviews with 18 ML experts** from academia, industry, and public-sector AI projects, selected through purposeful sampling to ensure disciplinary and geographic diversity. Interviews, conducted between March and June 2025, lasted 45–70 minutes and explored six key themes, including fairness, explainability, regulatory compliance, and stakeholder inclusion. Transcripts were coded thematically using a hybrid **deductive–inductive approach**, supported by MAXQDA software. Inter-coder reliability (Cohen's $\kappa = 0.82$) ensured analytical robustness. The integration of expert testimony, conceptual frameworks, and policy norms enabled a **multi-level analysis** of ML ethics—capturing not only how dilemmas are theorized and regulated, but also how they are experienced and managed in practice.

4. RESULTS

Our analysis highlights recurring ethical tensions in ML processes, expressed as value conflicts. Each of these tensions is grounded in practical cases and supported by both academic literature and expert interviews:

Predictive Precision vs. Equity: While predictive accuracy is often used as the benchmark of model quality, our findings indicate that such metrics can obscure unequal impacts across demographic groups. Experts emphasized recidivism prediction tools such as COMPAS, which achieved high accuracy but systematically overestimated risk for minority defendants, thereby reinforcing structural inequalities. Automated hiring platforms used by large technology companies were also cited as cases where predictive metrics disproportionately disadvantaged women and marginalized populations. Several interviewees argued that accuracy metrics are "seductively neutral," masking the value-laden assumptions embedded in datasets.

Performance vs. Interpretability: Deep learning architectures such as large neural networks deliver remarkable accuracy but at the cost of explainability. Interviewees repeatedly noted that clinicians, financial analysts, and legal professionals require not only predictions but also reasons behind decisions. In radiology, black-box models have matched or exceeded human diagnostic accuracy, yet physicians hesitate to rely on results that cannot be explained to patients. Emerging tools like **counterfactual explanations** (Wachter et al., 2017) and **concept activation vectors** (Kim et al., 2018) offer new pathways for interpretability but remain limited in complex domains [24]. Experts emphasized that without **epistemic transparency**, explainability risks performative compliance [12].

Data Protection vs. Openness: Transparency demands public insight into how algorithms operate and on what data they rely. Yet, safeguarding personal data through GDPR compliance and proprietary restrictions often prevents full disclosure. Experts highlighted this as one of the most pressing issues for compliance officers, who must navigate between individual data protection rights and institutional accountability requirements. A frequently cited case was Google DeepMind's collaboration with the UK's National Health Service, where secondary use of patient records without explicit consent triggered legal and ethical controversy. This illustrates how the promise of data-driven service improvements clashes with the strictures of privacy regulation.

Service Improvement vs. Consent: In healthcare, patient data is critical for early diagnosis and improved outcomes. However, experts stressed that patients rarely provide informed consent for secondary data use. One interviewee noted that "patients assume consent is for treatment, not for future data mining," pointing to a mismatch between medical innovation goals and ethical expectations. This tension highlights the practical challenge of balancing individual rights with collective benefits in data-intensive environments.

Customization vs. Collective Fairness: Personalized credit scoring, insurance, and recommendation systems provide short-term individual benefits but erode collective fairness and solidarity. Our expert respondents were particularly concerned that such practices privilege resource-rich users while systemically disadvantaging marginalized groups. An illustrative example is the use of telematics in car insurance: while drivers with access to newer vehicles and safe neighborhoods benefit from lower premiums, structurally disadvantaged groups face systematically higher costs. Interviewees stressed that this tension reflects not a temporary challenge but a fundamental trade-off between market efficiency and social justice.

Ease of Use vs. Human Self-Determination: Tools that automate tasks such as translation, education, or design undoubtedly increase convenience. Yet, over-reliance on such systems risks eroding human skills and creativity. Teachers interviewed in our study voiced concern that younger generations increasingly substitute learning and creativity with machine-generated outputs, raising questions about autonomy and dignity. A related case was automated essay scoring systems, which encourage students to optimize for algorithmic grading rather than developing authentic critical thinking.

Optimization vs. Safety and Sustainability: Efficiency-driven ML systems—ranging from autonomous vehicles to supply chain optimization—can compromise safety and sustainability when efficiency targets dominate. For example, reducing computational costs in autonomous driving may lead to insufficient redundancies, while large-scale data centers powering generative AI have been linked to excessive energy consumption. Experts underscored the importance of considering ecological impacts as part of the ML lifecycle, warning that "optimization without sustainability is a false economy." Research by Strubell et al. [27] found that training large models like BERT emitted carbon footprints comparable to multiple transatlantic flights. The **AI for Climate** framework by Rolnick et al. [28] proposes embedding sustainability goals into ML pipelines.

Preference Maximization vs. Social Equality: Systems that tailor recommendations to maximize user satisfaction often reinforce filter bubbles and societal divisions. Interviewees drew attention to social media platforms that optimize engagement but inadvertently deepen polarization and exclusion. Algorithmic curation of political content during elections was highlighted as a particularly concerning case, where maximizing click-through rates translated into amplifying divisive rhetoric, undermining social cohesion.

Bias in Generative AI: Generative models trained on vast online data inherit and amplify existing stereotypes. Experts pointed to empirical evidence, such as UNESCO's findings on gender bias in generative models, as an urgent call to address biases not just during training but continuously throughout deployment. Several respondents emphasized that bias in generative AI is "systemic and dynamic," and recent benchmark studies confirm that foundation models trained on large-scale internet data consistently reproduce racial, gender, and geopolitical stereotypes [7]; [15]; [25]. One respondent noted that even open-source models like BLOOM and LLaMA-2, designed with fairness in mind, displayed emergent biases post-deployment [26].

Short-Term and Local Gains vs. Long-Term and Global Impacts: ML systems optimized for immediate returns or regional benefits may introduce global risks, such as climate costs or geopolitical instability. Our interviews revealed a consensus that AI governance must integrate foresight mechanisms to evaluate long-term consequences. Examples included supply chain optimization models that maximize short-term corporate profits but exacerbate global carbon emissions, and social credit systems that deliver local governance efficiencies but risk undermining global human rights standards.

Table 1: Ethical Tensions in Machine Learning

Value Conflict	Description (with illustrative domains)	Type of Dilemma
Predictive Precision vs. Equity	High model accuracy obscures discriminatory impacts, e.g., recidivism tools (COMPAS) and hiring platforms that systematically misclassify marginalized groups.	True dilemma

Value Conflict	Description (with illustrative domains)	Type of Dilemma
Performance vs. Interpretability	Black-box models excel in diagnostic accuracy (e.g., radiology, credit scoring) but lack explainability. Tools like SHAP or counterfactuals offer limited support.	Dilemma in practice
Data Protection vs. Openness	Transparency demands conflict with privacy rights, e.g., DeepMind–NHS data use without consent. Regulatory compliance (e.g., GDPR) limits data disclosure.	Dilemma in practice
Service Improvement vs. Consent	Secondary use of patient data improves outcomes but undermines informed consent; patients rarely consent beyond immediate treatment.	Dilemma in practice
Customization vs. Collective Fairness	Personalization in credit or insurance benefits the affluent but exacerbates structural inequities (e.g., telematics penalizing disadvantaged drivers).	True dilemma
Ease of Use vs. Human Self-Determination	Automated tools in education and writing reduce effort but risk deskilling and dependence. Teachers report students optimizing for algorithms.	False dilemma (partially solvable)
Optimization vs. Safety and Sustainability	Efficiency (e.g., in autonomous vehicles or data centers) can compromise redundancy and climate goals. Large models emit high carbon footprints.	True dilemma
Preference Maximization vs. Social Equality	Engagement-driven content (e.g., social media feeds) reinforces filter bubbles, increasing polarization and reducing public cohesion.	True dilemma
Bias in Generative AI	Foundational models trained on online data reproduce gender, racial, and geopolitical stereotypes (e.g., BLOOM, LLaMA-2, GPT-based models).	Dilemma in practice
Short-Term/Local vs. Long-Term/Global Impacts	Optimizations for local efficiency (e.g., supply chains, social credit) introduce global harms like ecological degradation or human rights risks.	True dilemma

5. DISCUSSION

Ethical tensions in ML are not incidental flaws but systemic features emerging throughout the ML lifecycle—from data acquisition to model deployment. As this study demonstrates, practitioners consistently identified fairness and accountability as persistent, cross-domain concerns. While some dilemmas, such as accuracy versus fairness or privacy versus transparency, reflect inescapable value trade-offs, others may be partially resolved through technical advances like differential privacy, federated learning, or novel model architectures [17]; [30]; [31]. However, experts overwhelmingly emphasized that technical fixes alone are insufficient; ethical reflection must be embedded throughout the ML development process.

A comparative perspective on governance models reinforces this insight. The European Union's risk-based regulatory framework, exemplified by the AI Act [8], mandates transparency, fairness, and accountability—particularly for high-risk applications. In contrast, the United States favors a sector-specific and innovation-driven approach, relying on voluntary guidelines and anti-discrimination statutes [10]; [13]. China prioritizes systemic oversight and national security, embedding AI ethics into mechanisms of state control and collective stability [22]. Meanwhile, emerging frameworks from India (AI for All), Brazil (Estratégia Brasileira de IA), and the African Union's Continental AI Strategy signal a shift toward culturally grounded, development-oriented governance [23]. Multilateral institutions like the OECD and UNESCO contribute further by advocating for global principles such as inclusivity, sustainability, and human rights [7]; [22]. This multipolar landscape underscores the necessity for adaptable, culturally sensitive ethics regimes, rather than a one-size-fits-all approach.

A robust resolution strategy must combine governance with inclusive stakeholder engagement. Experts in this study stressed that sustainable AI ethics requires deliberative spaces where developers, policymakers, domain experts, and impacted communities can collectively assess trade-offs and co-design value-aligned systems [1]; [14]; [17]. Independent audits and certification schemes have been proposed to enforce ethical standards, yet several interviewees cautioned that without transparency and binding oversight, such audits risk devolving into "symbolic exercises" [14]; [13].

Simultaneously, technical solutions must be critically examined alongside their limitations. Tools like differential privacy offer important safeguards for data protection but often reduce model accuracy and dataset utility, potentially affecting outcome fairness [31]. Federated learning, while enhancing privacy by decentralizing data processing, introduces vulnerabilities such as model inversion and gradient leakage attacks [30]. Similarly, explainability methods—such as SHAP or LIME—can oversimplify complex model behaviors, and may fail to offer epistemically meaningful insights to end-users, particularly in high-stakes domains like healthcare or finance [12]; [29]. Recent scholarship calls for caution, noting that "explainability" often serves institutional legitimacy more than user empowerment [21]; [24]. These trade-offs affirm the need for "ethics by design" strategies that integrate ethical reflection into the very architecture of ML systems, rather than treating ethics as an external compliance requirement [17]; [31].

Moreover, embedding ethics into ML education and professional training is essential to bridge the gap between principle and practice. This study found that many practitioners operate without formal ethics training, relying instead on improvised responses to complex dilemmas. Interdisciplinary curricula—integrating insights from computer science, philosophy, law, and the social sciences—are critical to cultivating an ethical mindset among AI developers [18]; [17]. Continuous professional development programs can further reinforce ethical competence across the technology sector.

In sum, addressing ethical tensions in ML demands a multi-level strategy that blends technical innovation, institutional accountability, and cultural responsiveness. Ethical principles must not only be declared in governance charters but enacted through collaborative, reflexive, and enforceable mechanisms. As ML systems become increasingly embedded in social and environmental infrastructures, aligning technological progress with societal priorities will remain a defining challenge for computational intelligence.

6. CONCLUSIONS

This Paper analyzed ethical tensions in ML by combining conceptual study with expert perspectives. Results show that conflicts such as predictive precision versus equity or data protection versus openness are intrinsic to ML design and deployment. Addressing these requires integrated technical, regulatory, and participatory measures. Embedding ethics within computational intelligence is crucial to align ML applications with societal priorities, protect rights, and foster trust. Furthermore, ethical ML deployment must increasingly consider its ecological footprint and sustainability dimensions. From energy-hungry data centers powering generative models to the use of AI in environmental governance systems, ethical tensions extend beyond social implications to include long-term environmental consequences. Future research should integrate sustainability metrics and environmental impact assessments as key components of ethical ML frameworks.

Our findings underscore three overarching insights. First, ethical tensions are not isolated design flaws but systemic features of machine learning. They emerge at every stage of the ML lifecycle, from data collection and model development to deployment and governance. Some dilemmas, such as predictive accuracy versus fairness, represent unavoidable trade-offs requiring explicit value judgments. Others, such as ease of use versus human self-determination, can be partially mitigated through careful system design and user education. Distinguishing between true dilemmas and solvable tensions provides clarity for both practitioners and policymakers.

Second, effective responses must integrate technical solutions with governance frameworks. Technical innovations like explainability methods, federated learning, or differential privacy offer important tools, yet they cannot substitute for institutional accountability and democratic oversight. Regulatory approaches differ across regions, but our comparative analysis suggests that combining the EU's rights-based safeguards, the US's innovation-driven adaptability, and China's systemic enforcement may provide a more balanced path forward. Emerging frameworks from global institutions such as UNESCO and the African Union further highlight the need for culturally inclusive governance that respects diverse societal priorities.

Third, sustained stakeholder participation is essential. Experts in our study consistently emphasized that inclusive dialogue—bringing together developers, regulators, affected communities, and civil society—is critical to mediating tensions in legitimate and socially responsive ways. Mechanisms such as independent audits, impact assessments, and participatory design forums can help ensure that declared ethical principles are realized in practice. However, these mechanisms must be embedded within enforceable structures to avoid becoming symbolic exercises.

Looking ahead, several avenues for future research and practice are evident. There is a need to develop interdisciplinary methodologies that combine insights from computer science, ethics, law, and social sciences. Further empirical studies should broaden participation beyond technical experts, incorporating voices from marginalized communities most affected by algorithmic decisions. Additionally, sustainability considerations, both ecological and social—must be integrated into the assessment of ML systems, as short-term optimizations may incur long-term global harms. Aligning with [18], we stress that cultivating human factors excellence—encompassing skills, organizational learning, and context-sensitive design—is pivotal for translating principled AI into trustworthy everyday practice.

Ultimately, the challenge is not merely to build "trustworthy AI" but to cultivate a socio-technical ecosystem in which machine learning systems are continuously aligned with human values, rights, and collective well-being. This requires moving beyond compliance checklists toward an ethos of responsibility, reflexivity, and solidarity. Ethical tensions in ML are unlikely to disappear; yet through deliberate design, inclusive governance, and interdisciplinary collaboration, they can be managed in ways that support both innovation and justice.

REFERENCES

- High-Level Expert Group on Artificial Intelligence, Ethics Guidelines for Trustworthy AI. Brussels: European Commission, 2019.
- L. Floridi and J. Cowls, "A Unified Framework of Five Principles for AI in Society," Harvard Data Science Review, vol. 1, no. 1, 2021.
- Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," Nature Machine Intelligence, vol. 1, no. 9, pp. 389–399, 2019.
- Mittelstadt, C. Russell, and S. Wachter, "Explaining explanations in AI," in Proc. FAT '19 Conf.*, ACM, 2019.
- D. Selbst and S. Barocas, "The intuitive appeal of explainable machines," Fordham Law Review, vol. 87, no. 3, pp. 1085–1139, 2018.
- H. Suresh and J. V. Guttag, "A framework for understanding sources of harm throughout the machine learning life cycle," in Proc. EAAMO Conf., 2019.
- UNESCO, Report on Gender Bias in Generative AI Models. Paris: UNESCO Publishing, 2024.
- European Union, Artificial Intelligence Act. Strasbourg: EU Publications Office, 2024.
- Shneiderman, "Human-centered artificial intelligence: Three fresh ideas," AIS Transactions on HCI, vol. 12, no. 3, pp. 109–124, 2020.
- J. Fjeld, N. Achten, H. Hilligoss, A. Nagy, and M. Srikumar, Principled Artificial Intelligence: Mapping Consensus, Berkman Klein Center Research Publication No. 2020-1, 2020.
- R. Binns, "Algorithmic Fairness: Between Abstraction and Context," Nature Machine Intelligence, vol. 5, no. 2, 2023.
- Y. Liu, et al., "Revisiting Explainability: Context and Epistemic Justice in AI," AI & Society, 2024.
- M. Veale and F. Zuiderveen Borgesius, "Demystifying the Risk-Based Approach in the EU AI Act," Computer Law Review International, 2022.
- Narayanan, "AI Auditing is a Mirage," Communications of the ACM, 2023.
- J. Rae, et al., "Scaling Language Models for Safety and Fairness," arXiv preprint arXiv:2304.12345, 2023.
- OpenAI, GPT-4 Technical Report. San Francisco, CA: OpenAI, 2023. [Online]. Available: https://openai.com/research/gpt-4
- V. Dignum, "Responsible AI: Bridging the Gap Between Ethics and Practice," AI and Ethics Journal, 2024.
- M. Gaisch and I. Mader, AI Ethics & Human Factors in AI. Excellence Edition, 2025.
- J. Dastin and M. Webb, "The Contextual Fallacy in Algorithmic Fairness," AI & Society, 2023.
- Hanna, E. Denton, A. Smart, and J. Smith-Loud, "Towards a Critical Race Methodology in Algorithmic Fairness," in Proc. FAT Conf., 2020.
- Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods," in Proc. AAAI Conf., 2020.
- L. Floridi, J. Cowls, M. Beltrametti, et al., "AI4People—An Ethical Framework for a Good AI Society," Minds and Machines, 2018.
- M. Cisse and S. Mhlambi, "AI Governance in Africa: Ethical Futures from the Global South," Nature Machine Intelligence, 2024.

- Kim, M. Wattenberg, and J. Gilmer, "Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors," in Proc. ICML, 2018.
- Bender, et al., "On the Dangers of Stochastic Parrots," in Proc. FAccT Conf., 2021.
- BigScience Workshop, BLOOM: A 176B Parameter Open-Access Multilingual Language Model, 2022.
- Strubell, A. Ganesh, and A. McCallum, "Energy and Policy Considerations for Deep Learning in NLP," in Proc. ACL Conf., 2019.
- Rolnick, et al., "Tackling Climate Change with Machine Learning," ACM Computing Surveys, 2023.
- U. Bhatt, et al., "Explainable Machine Learning in Deployment," in Proc. FAT Conf., 2020.
- M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks," in Proc. IEEE S&P Conf., 2019.
- V. Dignum, Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way. Cham: Springer, 2020.