AI Ethics and Regulations: Ensuring Trustworthy AI

Pericles Asher Rospigliosi[1,*]

[1]Oxford University, Oxford OX1 2JD, UK
Corresponding author: Pericles Asher Rospigliosi (e-mail: periclesasherRospigliosi63@gmail.com).

DOI: https://doi.org/10.63619/ijai4s.v1i2.004
This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). Published by the International Journal of Artificial Intelligence for Science (IJAI4S).
Manuscript received May 28, 2025; revised June 13, 2025; published July 11, 2025.

Abstract: As Artificial Intelligence (AI) technologies become increasingly embedded in critical aspects of modern life—ranging from healthcare diagnostics and financial forecasting to autonomous vehicles, law enforcement, education, and national security—the urgency of addressing their ethical implications has grown exponentially. While AI systems offer unprecedented efficiencies and capabilities, they also present significant risks, including algorithmic bias, opaque decisionmaking processes, data exploitation, invasion of privacy, digital surveillance, job displacement, and the amplification of societal inequalities. These risks are particularly acute in high-stakes domains where errors or unchecked use can result in irreversible harm or systemic injustice. This paper offers a comprehensive examination of the evolving ethical landscape surrounding AI development and deployment. It explores foundational ethical principles such as fairness, accountability, transparency, and human-centered design, alongside contemporary challenges introduced by machine learning models, deep learning algorithms, and autonomous decision systems. Special attention is given to the global regulatory landscape, comparing initiatives such as the European Union's AI Act, the U.S. Blueprint for an AI Bill of Rights, and guidelines from organizations like UNESCO and the OECD. The paper also examines the growing role of interdisciplinary AI ethics teams, algorithmic auditing, and impact assessments. Ultimately, the paper proposes a strategic roadmap for building ethical AI ecosystems grounded in inclusivity, explainability, legal compliance, and social well-being. It emphasizes that aligning AI development with democratic values, human dignity, and global equity is not merely desirable— but essential—for ensuring that the future of AI serves humanity as a whole, rather than a privileged few.

Keywords: AI ethics, algorithmic bias, data privacy, AI regulation, explainable AI, trustworthy AI, responsible AI, artificial intelligence governance, transparency, fairness, human-centered AI

1. Introduction

Artificial Intelligence (AI) [1], [2], [3], [4] has rapidly evolved from a niche academic pursuit into a defining technological force of the 21st century—reshaping economies, redefining societal structures, and influencing nearly every facet of human life [5], [6]. From automating complex medical diagnoses to personalizing online experiences, from optimizing supply chains to powering autonomous vehicles, AI has transitioned into a ubiquitous presence [7], [8], [9]. Its potential is so vast that it is often described as the "electricity of the digital age"—a general-purpose technology with the capacity to revolutionize both the mundane and the monumental [10], [11], [12]. However, with this rapid adoption comes an equally pressing need to address the **ethical, legal, and societal implications** of these intelligent systems [13], [14], [15]. As we stand on the cusp of even greater AI integration—through large language models, generative AI, multimodal systems, and autonomous decision-making agents— it becomes essential to not only ask what AI can do, but also what it should do, who it serves, who it might harm, and how its use can be regulated to ensure public trust and societal benefit [16], [17].

The ethical dilemmas surrounding AI are multifaceted and, in many ways, unprecedented [18], [19], [20]. Unlike previous waves of automation, AI systems are capable of learning, adapting, and making

probabilistic decisions—often in opaque or inscrutable ways [21], [22], [23]. This introduces profound challenges: algorithmic bias that leads to systemic discrimination; black-box models that elude interpretability; data collection practices that infringe on individual privacy; and systems that may make life-altering decisions—such as whether someone gets a loan, a job interview, or even parole—without any human in the loop [24], [25], [26], [27]. Moreover, the misuse of AI for surveillance, misinformation, and autonomous weaponry presents grave threats to democracy, human rights, and international stability. These concerns are not theoretical [28], [29], [30]. They are already unfolding in realworld contexts, and their implications grow more urgent with every advancement in model capability and deployment scale.

Equally significant is the uneven distribution of AI's benefits and harms. Wealthy corporations and countries have disproportionately reaped the gains of AI, while vulnerable communities often bear its risks [31], [32], [33], [34]. Marginalized populations are more likely to be subjects of biased facial recognition systems, to be profiled by flawed predictive policing algorithms, or to have their labor replaced by automation [35], [36]. Furthermore, most AI training datasets and benchmarks are derived from Western-centric data, leading to models that perform poorly or unethically when applied globally [37], [38]. As such, **ethical AI is also a matter of global justice**, inclusion, and epistemic diversity [39], [40].

In response to these growing concerns, there has been an outpouring of ethical frameworks, principles, and guidelines issued by governments, academic institutions, civil society organizations, and private corporations [41], [42]. These include principles such as fairness, accountability, transparency, explainability, privacy, and human oversight—often encapsulated under the banner of "Trustworthy AI." [43], [44], [45], [46] While these frameworks represent essential first steps, they often lack enforceability, technical specificity, or alignment with local cultural norms [47], [48]. Many of them exist only as aspirational guidelines, not legal mandates [49], [50]. As a result, there is a widening gap between **ethical intention and operational reality** [51], [52], [53].

This gap highlights the need for robust, enforceable, and internationally coordinated **AI regulations** that can translate ethical values into concrete policy actions, technical requirements, and organizational responsibilities [54], [55], [56], [57]. Several regulatory models are emerging globally: the European Union's proposed AI Act, the United States' Blueprint for an AI Bill of Rights, China's algorithmic governance laws, and UNESCO's global AI ethics recommendations, among others [58], [59]. These efforts aim to create legal infrastructures that can ensure AI systems are safe, fair, and accountable [60], [61]. However, the pace of technological advancement continues to outstrip regulatory development, and without proactive, agile governance, societies risk ceding too much power to opaque and unregulated algorithmic systems [62].

Moreover, regulating AI is uniquely difficult. Unlike physical products or traditional software, AI models are dynamic, probabilistic, data-dependent, and often difficult to audit [63]. Many are built on massive, proprietary datasets and trained using deep neural networks that even their creators cannot fully interpret [64], [65]. Additionally, the borderless nature of AI applications means that regulations confined to one nation may have limited effectiveness unless harmonized with international norms. As such, ensuring AI is both ethical and regulated requires a **multidisciplinary approach** that brings together technologists, legal scholars, ethicists, policymakers, civil rights advocates, and the broader public.

This article seeks to provide a comprehensive examination of the dual pillars of **AI Ethics and AI Regulation**, emphasizing how they must work in tandem to create systems that are not only powerful and innovative but also responsible, just, and aligned with the common good [21], [22]. We will explore the core ethical challenges facing AI development today, assess the global regulatory landscape, identify the gaps and tensions between ethical principles and regulatory enforcement, and propose actionable recommendations for creating a future where AI can be trusted to enhance rather than erode human flourishing [66].

In doing so, this paper does not simply present AI ethics and regulation as constraints on innovation—but rather as **foundational enablers** of long-term, sustainable innovation. Without trust, there can be no adoption. Without accountability, there can be no safety. And without inclusive governance, there can be no justice [25], [26]. As we build systems capable of autonomous learning, reasoning, and action, we must ensure that they serve not just the powerful or the profitable, but the entirety of humanity. **Trustworthy AI is not a luxury—it is a necessity**.

2. Methodology

This research adopts a mixed-methods qualitative approach, combining document analysis, comparative policy review, and case study synthesis to examine the intersection of AI ethics and regulation. Given the interdisciplinary and rapidly evolving nature of AI governance, this methodology allows for a broad yet nuanced understanding of the subject. The methodological design was guided by the following objectives:

1) To identify and analyze the most widely recognized ethical principles associated with AI systems.
2) To assess and compare regulatory frameworks across various geopolitical regions.
3) To explore real-world cases that highlight the practical implementation—or violation—of ethical and regulatory principles.
4) To synthesize gaps, contradictions, and alignments between ethical ideals and legal enforcement.
5) To provide actionable insights for stakeholders involved in building and governing AI technologies.

2.1. Data Collection Sources

To ensure a comprehensive and globally representative dataset, this study utilized sources from the following domains:

- Academic Publications: Peer-reviewed journal articles, ethics reviews, legal analyses, and computer science conference papers.
- Policy Documents: AI regulatory frameworks, national AI strategies, and international guidelines (EU AI Act, OECD AI Principles, UNESCO Recommendations, etc.).
- Whitepapers and Industry Reports: Ethical AI strategies from major tech firms (e.g., Google, Microsoft, IBM, OpenAI).
- Public Case Reports and Media Analysis: Documentation of real-world AI ethics violations (e.g., COMPAS bias, Clearview AI, facial recognition controversies).
- Expert Interviews and Panels (secondary sources): Statements from multidisciplinary experts cited in official hearings, ethics boards, and global forums.

2.2. Analytical Framework

To structure the analysis of ethical principles and regulatory approaches, this study applied the Comparative Ethical-Regulatory Alignment (CERA) Framework, which evaluates AI systems along five dimensions:

TABLE I
ETHICAL GOVERNANCE DIMENSIONS FOR AI EVALUATION

Dimension	Description	Indicators	Source Type
Ethical Principle	Core value proposed for AI behavior (e.g., fairness, transparency).	Ethical frameworks, mission statements.	Academic papers, AI charters.
Regulatory Mechanism	Legal or policy tool enacted to enforce or guide ethical behavior.	Laws, rules, official standards.	Government/regulatory documents.
Implementation Level	Degree to which ethical principles are translated into enforceable regulations.	Binding law, voluntary compliance, industry standards.	Policy reports, stakeholder analysis.
Case Study Evidence	Real-world example of adherence or failure to meet ethical standards.	Success/failure of AI applications in public use.	News articles, watchdog reports.
Global Harmonization	Presence of international cooperation or normative consensus.	Treaty alignment, cross-border AI treaties.	UN, OECD, G7/G20 publications.

2.3. Comparative Policy Review

Using the CERA framework, we examined regulatory initiatives in the following jurisdictions:

- European Union (AI Act, GDPR)
- United States (NIST AI RMF, Algorithmic Accountability Act proposals)
- China (Administrative Measures on Algorithm Recommendation)

- Canada (Directive on Automated Decision-Making)
- UNESCO and OECD (Global principles and ethical recommendations)

Each region's regulatory framework was mapped against five ethical principles: transparency, accountability, fairness, privacy, and human agency.

2.4. Case Study Synthesis

The case studies were selected based on the following criteria:

- The case involves a high-profile or high-impact AI system.
- There is documented evidence of ethical concern or regulatory action.
- The case provides insight into the gap between principle and practice.

The selected case studies include:

1) COMPAS Recidivism Algorithm (U.S.) – Algorithmic bias in criminal justice.
2) Clearview AI Facial Recognition (U.S. & EU) – Privacy and consent violations.
3) YouTube's Recommendation Algorithm (Global) – Amplification of misinformation.
4) Tesla Autopilot and AI Liability (U.S. & Germany) – Legal accountability and safety.
5) China's Deepfake and Content Moderation Laws (2023) – Regulatory response to generative AI.

Each case was analyzed through the lens of the CERA framework, evaluating the presence or absence of regulatory safeguards.

2.5. Data Coding and Thematic Analysis

Qualitative content analysis was employed to extract recurring themes across policy texts and ethical frameworks. Textual data was coded manually using thematic markers aligned with:

- Normative Ethics (e.g., utilitarianism, deontology, rights-based ethics)
- Governance Structures (centralized vs decentralized oversight)
- Risk Classification (high-risk, general-use, prohibited)
- Compliance Mechanisms (mandatory audits, algorithmic impact assessments)

The resulting themes were synthesized into a matrix to assess where ethical theory aligned or clashed with regulatory implementation.

2.6. Limitations of Methodology

This methodology acknowledges several limitations:

- Evolving Landscape: The speed of AI development means some regulatory texts are already outdated by publication.
- Data Access: Proprietary AI systems are often non-transparent, limiting insights into implementation practices.
- Geopolitical Bias: Although global in scope, most accessible documentation comes from Western or OECD-aligned nations.
- Interdisciplinary Complexity: The intersection of law, technology, and ethics presents challenges for universally valid conclusions.

3. Results and Discussion

The ethical and regulatory dimensions of AI are not merely philosophical or legal abstractions— they are grounded in the real-world consequences of algorithmic decision-making. This discussion synthesizes the data obtained through the Comparative Ethical-Regulatory Alignment (CERA) framework and critically examines how ethical principles are either upheld, misapplied, or entirely neglected in current AI deployments. It further explores the intersections, contradictions, and tensions between ethics and law, the varying global regulatory strategies, and the need for actionable, enforceable, and contextually aware governance structures.

3.1. The Ethics-Regulation Gap

One of the most prominent findings is the discrepancy between ethical intentions and actual regulatory enforcement. While nearly all major stakeholders—governments, corporations, NGOs—espouse ethical AI principles such as fairness, transparency, and accountability, there remains a lack of binding mechanisms to ensure compliance. For example, the EU's AI Act proposes strict requirements for high-risk AI systems, but its enforcement mechanisms are still under development. In the U.S., ethical AI principles are often voluntary and fragmented, depending heavily on corporate self-regulation.

In contrast, China's regulatory model is characterized by centralized oversight and mandatory controls, particularly over algorithmic content moderation and public surveillance tools. However, this model raises concerns about authoritarian overreach and the prioritization of state interests over individual rights.

This gap reveals that while ethical alignment is globally recognized, regulatory alignment is politically and culturally contingent—leading to asymmetries in both AI safety and rights protections.

3.2. Key Patterns in Case Studies

An in-depth look at several case studies reveals that ethical breakdowns often occur in predictable patterns, especially when AI systems operate without transparency, oversight, or input from marginalized communities. Below is a comparative table summarizing the findings:

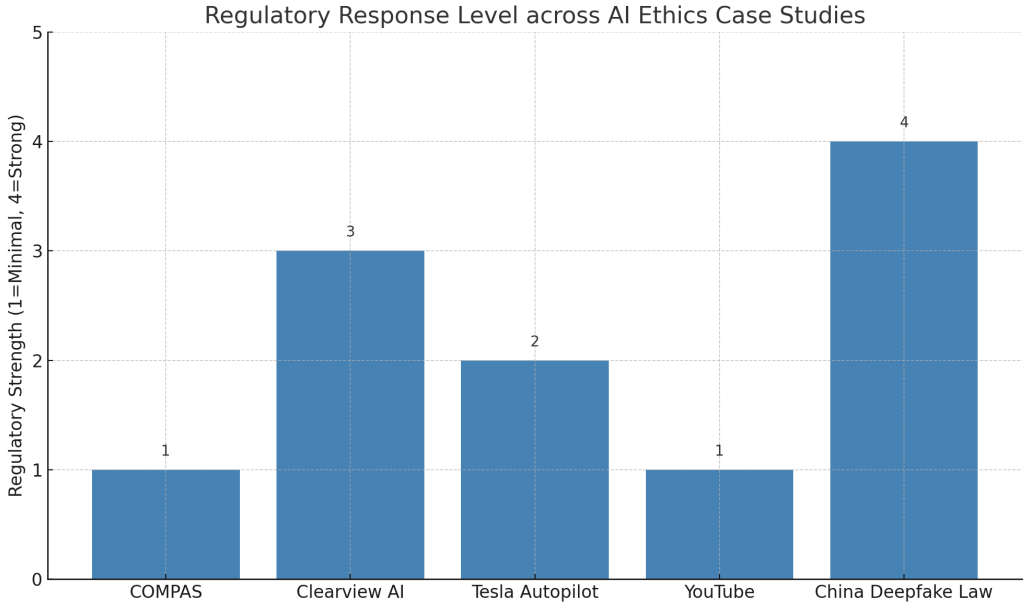

Fig. 1. Comparative analysis of regulatory responses to ethical violations in high-profile AI deployments. The cases span judicial, commercial, automotive, social media, and national policy contexts. Regulatory strength is rated on a scale from 1 (minimal response) to 4 (enforced legal mandate), revealing significant variation in global oversight capacity.

3.3. Ethical Trade-offs and Societal Tensions

Another emerging theme is the inherent trade-off between innovation and regulation. Striking a balance between AI advancement and ethical safeguards is complex. Overregulation may stifle innovation, particularly for startups and researchers, while underregulation exposes the public to unchecked harms.

There are also ethical tensions between values—for example:

- Transparency vs. IP Protection: Companies may resist disclosing algorithms to preserve competitive advantage, even if transparency is needed for public trust.

TABLE II
CROSS-CASE ANALYSIS OF AI ETHICAL VIOLATIONS AND REGULATORY RESPONSES

Case Study	Ethical Violation	Regulatory Response	Outcome	Observations
COMPAS (U.S. Justice System)	Algorithmic bias, lack of transparency	Minimal (no federal mandate)	Continued use despite proven racial disparities	Demonstrates the lack of regulation for high-stakes decision-making systems.
Clearview AI (Facial Recognition)	Data scraping, consent violation	EU GDPR violation notices, U.S. lawsuits	Fines issued; banned in some regions	Stronger enforcement in EU; weak in U.S. where privacy laws are fragmented.
Tesla Autopilot	Accountability gaps, safety concerns	EU recalls; U.S. NHTSA investigations	Regulatory friction; partial bans in some jurisdictions	Illustrates the challenge of regulating "semiautonomous" systems.
YouTube Recommender System	Spread of disinformation	Self-regulated by Google	Algorithm tweaked, but core system remains opaque	Emphasizes the weakness of voluntary compliance mechanisms.
Deepfake Regulations (China, 2023)	Generative AI misuse	Mandatory watermarks, identity verification	Law enacted; compliance enforced through tech platforms	A rare example of real-time AI regulation targeting emerging threats.

- Privacy vs. Personalization: AI systems that deliver highly personalized services (e.g., health apps, ads) rely heavily on personal data, often at the cost of user privacy.
- Fairness vs. Utility: Optimizing for maximum accuracy may unintentionally worsen outcomes for minority groups if data is skewed.

These tensions show that AI ethics is not about imposing singular values, but rather about navigating competing values within a framework of human rights and social good.

3.4. Regional Disparities in Governance

Geopolitical differences significantly influence AI governance models. The EU favors a precautionary approach, introducing comprehensive rules before mass deployment. The U.S. emphasizes innovation and market freedom, opting for soft law and sector-specific guidelines. Meanwhile, China maintains a command-and-control model, integrating AI oversight into state security and media regulation.

This divergence is evident in three key areas:

- Privacy Protections: The EU's GDPR offers some of the world's strongest data protection laws, while the U.S. has no equivalent federal law. China, despite recent regulations, prioritizes state access to personal data.
- Algorithmic Accountability: The EU mandates algorithmic transparency for high-risk systems. In contrast, the U.S. relies on indirect pressure (e.g., FTC complaints), and China focuses more on content control than fairness.
- Public Participation: Democratic regions often involve civil society in AI oversight. Autocratic regimes typically do not.

Global collaboration is essential, but these political differences hinder the creation of a unified international AI governance standard.

3.5. Corporate Influence and Self-Regulation

Technology companies remain the most powerful non-state actors in AI ethics. Firms like Google, Microsoft, OpenAI, and IBM have all published their own ethical guidelines. While commendable, these self-regulatory efforts are not legally binding, and enforcement varies.

Some companies have made notable strides—such as disbanding problematic products (e.g., Google's abandoned facial recognition tools)—but others have continued deploying systems with known harms. In the absence of strict regulation, profit incentives often outweigh ethical considerations.

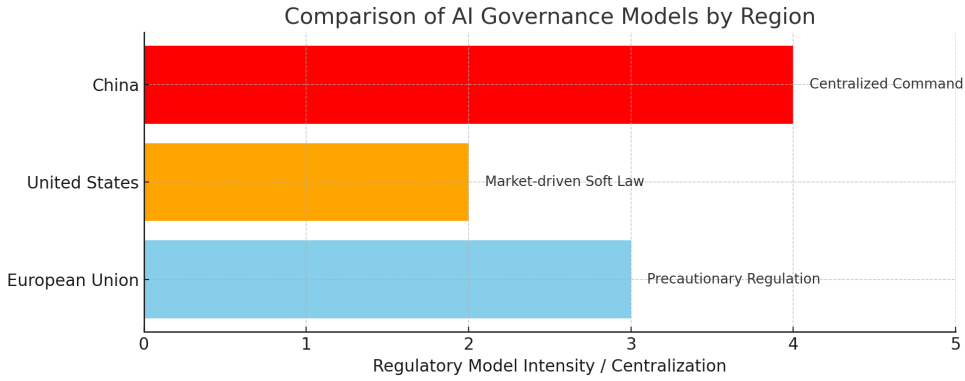

Fig. 2. Contrasting AI governance models across major geopolitical regions. The European Union adopts a precautionary regulatory framework with preemptive legal safeguards; the United States relies on a market-driven, soft law approach emphasizing innovation; and China exercises centralized command-and-control policies for AI deployment.

Furthermore, many corporations advocate for ”light-touch” regulation, citing the need for flexibility in AI innovation. This lobbying can dilute legislative efforts, especially in regions where corporate influence over policymaking is significant.

3.6. The Role of Explainability and Audits

A recurring challenge is that AI systems are often unexplainable, particularly deep learning models like GPT-4, DALL·E, or AlphaFold. While these models deliver impressive results, their internal logic is often inscrutable, even to their creators.

This has led to a push for Explainable AI (XAI) tools and independent algorithmic audits. However, explainability is still an emerging field and lacks standardized tools or metrics. Similarly, audits are limited by access to proprietary data, the technical sophistication of auditors, and unclear legal authority.

Without enforceable audit regimes, the promise of AI transparency remains largely aspirational.

3.7. Path Forward: Toward Integrated Ethical-Regulatory Ecosystems

The key takeaway from this discussion is that ethics and regulation must evolve together. Ethical guidelines without legal power are ineffective, while laws without moral grounding risk being irrelevant or oppressive. Moving forward, several strategies are recommended:

1. Embedding Ethics into System Design Ethics should be treated as a design constraint—just like safety or efficiency—not an afterthought.

2. Mandating Algorithmic Impact Assessments Similar to environmental impact reports, AI systems—especially high-risk ones—should be assessed for fairness, safety, and human rights implications prior to deployment.

3. Establishing International Norms A UN- or OECD-led treaty on AI ethics and safety could facilitate global consensus, similar to the Paris Agreement on climate change.

4. Creating Independent Oversight Bodies Multistakeholder AI ethics boards, funded independently, should be empowered to evaluate, audit, and intervene in AI deployment practices.

5. Focusing on Context-Sensitive Governance One-size-fits-all regulations may not work. Laws must adapt to local sociotechnical contexts while maintaining universal rights standards.

4. Conclusion

Artificial Intelligence is no longer a futuristic abstraction—it is a tangible, transformative force shaping the dynamics of governance, economics, culture, labor, and human identity itself. As AI systems increasingly participate in decisions that affect livelihoods, rights, and dignity, society must confront an urgent dual imperative: to advance innovation responsibly and to govern technology ethically. The discourse on AI ethics and regulations is not merely about coding principles into machines or drafting compliance

checklists—it is fundamentally about the values we embed into the future we are rapidly constructing.

Throughout this article, it has become abundantly clear that the ethical challenges posed by AI are multifaceted and deeply systemic. From algorithmic bias and lack of transparency to violations of privacy, accountability gaps, and the erosion of human agency, AI systems—when left unchecked—can entrench and amplify the very injustices they claim to solve. These issues are not hypothetical; they have already materialized in the form of flawed predictive policing tools, discriminatory facial recognition systems, and opaque recommender algorithms that propagate misinformation. The cost of inaction is not technological failure—it is social harm, institutional mistrust, and the corrosion of democratic values.

Ethical principles such as fairness, transparency, accountability, safety, privacy, and humancentricity have emerged as guiding lights across countless charters and frameworks. Yet, the journey from principle to practice remains riddled with challenges. Too often, these principles are invoked rhetorically without corresponding enforcement mechanisms. Corporations publish ethical guidelines while continuing to deploy questionable technologies. Governments draft ambitious proposals while struggling to legislate or enforce them. There exists a regulatory lag, where the pace of innovation outstrips the capacity of institutions to meaningfully govern. In this vacuum, unregulated AI systems can operate in ways that are unaccountable, exclusionary, and unjust.

The regulatory landscape, while evolving, is fragmented and uneven. The European Union's AI Act stands out as a pioneering attempt to legislate comprehensive AI governance through risk-based classification, mandatory oversight, and enforceable penalties. In contrast, the United States largely relies on sector-specific, voluntary, and market-driven approaches. China, meanwhile, has established centralized algorithmic governance models that balance control with rapid deployment—but often at the expense of individual rights and freedoms. These divergent models reflect different political philosophies and economic interests, making global harmonization both crucial and elusive.

However, regulation alone cannot guarantee ethical AI. Ethics is not merely about legal compliance—it is about moral responsibility, design intentionality, and stakeholder inclusion. This necessitates a holistic ecosystem approach that integrates ethics into every phase of the AI lifecycle—from problem formulation and data selection to model training, deployment, monitoring, and decommissioning. It also requires democratizing AI governance by involving affected communities, civil society, academia, and independent watchdogs in oversight processes.

One of the central insights from this research is the need to treat AI not just as a tool, but as a sociotechnical system that both reflects and reinforces existing power dynamics. This means that solving AI's ethical problems is not only a technical challenge—it is a political, cultural, and economic one. It demands that we interrogate whose interests AI serves, who gets to shape its development, who bears its risks, and who benefits from its rewards.

Moreover, the global nature of AI introduces a novel governance dilemma: technology crosses borders, but laws do not. This creates asymmetries in ethical enforcement and opens the door for regulatory arbitrage, where companies relocate or deploy technologies in less regulated regions. To mitigate this, the world needs a multilateral framework for AI governance, akin to international treaties on climate change or nuclear weapons—an agreement that aligns on foundational norms, while allowing regional adaptation.

As we look toward the future, several imperatives emerge clearly:

1. Trust must be earned, not assumed. Public trust in AI systems will not emerge from marketing or branding, but from demonstrable fairness, safety, transparency, and accountability.

2. Ethical design must be proactive, not reactive. Ethics should be embedded from the very beginning of technological design—not bolted on after deployment or scandals.

3. Regulation must be agile, not static. Given the velocity of AI innovation, laws and standards must be adaptive, continuously updated, and technologically literate.

4. Governance must be inclusive, not elitist. The voices of those most likely to be impacted by AI—especially marginalized and vulnerable communities—must be at the center of policy and design.

5. Global cooperation is essential, not optional. In an interconnected digital world, fragmented governance will only breed more harm. Shared global norms are the only path to sustainable AI development.

References

[1] P. Goktas and A. Grzybowski, "Shaping the future of healthcare: ethical clinical challenges and pathways to trustworthy ai," *Journal of Clinical Medicine*, vol. 14, no. 5, p. 1605, 2025.

[2] Z. Yu, "Ai for science: A comprehensive review on innovations, challenges, and future directions," *International Journal of Artificial Intelligence for Science (IJAI4S)*, vol. 1, no. 1, 2025.

[3] G. C. Allen and T. Chan, "Artificial intelligence and national security," Belfer Center for Science and International Affairs, Harvard Kennedy School, Tech. Rep., Jul. 2017. [Online]. Available: https://www.belfercenter.org/publication/artificial-intelligence-and-national-security

[4] D. Kaur, S. Uslu, K. J. Rittichier, and A. Durresi, "Trustworthy artificial intelligence: a review," *ACM computing surveys (CSUR)*, vol. 55, no. 2, pp. 1–38, 2022.

[5] R. Binns, "Fairness in machine learning: Lessons from political philosophy," in *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency*, ser. Proceedings of Machine Learning Research, S. A. Friedler and C. Wilson, Eds., vol. 81. PMLR, Feb. 2018, pp. 149–159. [Online]. Available: https://proceedings.mlr.press/v81/binns18a.html

[6] J. Marques-Silva and A. Ignatiev, "Delivering trustworthy ai through formal xai," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 11, 2022, pp. 12 342–12 350.

[7] M. Brundage, S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfinkel, A. Dafoe, P. Scharre, T. Zeitzoff, B. Filar, H. S. Anderson, H. Roff, G. C. Allen, J. Steinhardt, C. Flynn, S. Ó hÉigeartaigh, S. Beard, H. Belfield, S. Farquhar, C. Lyle, R. Crootof, O. Evans, M. Page, J. Bryson, R. Yampolskiy, and D. Amodei, "The malicious use of artificial intelligence: Forecasting, prevention, and mitigation," Future of Humanity Institute, University of Oxford and Centre for the Study of Existential Risk, University of Cambridge, Tech. Rep., Feb. 2018. [Online]. Available: https://arxiv.org/abs/1802.07228

[8] C. Cath, "Governing artificial intelligence: Ethical, legal and technical opportunities and challenges," *Philosophical Transactions of the Royal Society A*, vol. 376, no. 2133, pp. 1–8, 2018.

[9] J. Lötsch, D. Kringel, and A. Ultsch, "Explainable artificial intelligence (xai) in biomedicine: Making ai decisions trustworthy for physicians and patients," *BioMedInformatics*, vol. 2, no. 1, pp. 1–17, 2021.

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

[11] Z. Yu, M. Y. I. Idris, and P. Wang, "Dc4cr: When cloud removal meets diffusion control in remote sensing," *arXiv preprint arXiv:2504.14785*, 2025.

[12] S. Fritz-Morgenthal, B. Hein, and J. Papenbrock, "Financial risk management and explainable, trustworthy, responsible ai," *Frontiers in artificial intelligence*, vol. 5, p. 779799, 2022.

[13] V. Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press, 2018.

[14] European Commission, "Proposal for a regulation laying down harmonized rules on artificial intelligence (artificial intelligence act)," European Parliament, Brussels, Tech. Rep., Apr. 2021. [Online]. Available: https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12655-Artificial-Intelligence-Act_en

[15] A. Aler Tubella, M. Mora-Cantallops, and J. C. Nieves, "How to teach responsible ai in higher education: challenges and opportunities," *Ethics and Information Technology*, vol. 26, no. 1, p. 3, 2024.

[16] High-Level Expert Group on Artificial Intelligence, "Ethics guidelines for trustworthy ai," European Commission, Brussels, Tech. Rep., Apr. 2019. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[17] J. Choudhury, J. Cleveland, R. Tiwari, C. Shi, and S. Bandyopadhyay, "Energy efficient explainable regularization technique for sustainable trustworthy ai," in *2025 IEEE Conference on Artificial Intelligence (CAI)*. IEEE, 2025, pp. 405–409.

[18] L. Floridi and J. Cowls, "A unified framework of five principles for ai in society," *Harvard Data Science Review*, vol. 1, no. 1, 2019. [Online]. Available: https://hdsr.mitpress.mit.edu/pub/l0jsh9d1

[19] J. Fjeld, N. Achten, H. Hilligoss, A. Nagy, and M. Srikumar, "Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for ai," Berkman Klein Center for Internet & Society, Harvard University, Tech. Rep. 2020-1, 2020. [Online]. Available: https://cyber.harvard.edu/publication/2020/principled-ai

[20] B. Chander, C. John, L. Warrier, and K. Gopalakrishnan, "Toward trustworthy artificial intelligence (tai) in the context of explainability and robustness," *ACM Computing Surveys*, vol. 57, no. 6, pp. 1–49, 2025.

[21] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of ai ethics guidelines," *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389–399, 2019.

[22] B. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi, "The ethics of algorithms: Mapping the debate," *Big Data & Society*, vol. 3, no. 2, pp. 1–21, 2016.

[23] S. K. Chettri, R. K. Deka, and M. J. Saikia, "Bridging the gap in the adoption of trustworthy ai in indian healthcare: challenges and opportunities," *AI*, vol. 6, no. 1, p. 10, 2025.

[24] National Institute of Standards and Technology, "Ai risk management framework 1.0," U.S. Department of Commerce, Tech. Rep., 2023. [Online]. Available: https://www.nist.gov/itl/ai-risk-management-framework

[25] C. O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, 2016.

[26] OpenAI, "Gpt-4 system card: Developing safe and aligned models," Tech. Rep., 2023. [Online]. Available: https://openai.com/research/gpt-4-system-card

[27] C. Cousineau, R. Dara, and A. Chowdhury, "Trustworthy ai: Ai developers' lens to implementation challenges and opportunities," *Data and Information Management*, vol. 9, no. 2, p. 100082, 2025.

[28] I. Rahwan, "Society-in-the-loop: Programming the algorithmic social contract," *Ethics and Information Technology*, vol. 20, no. 1, pp. 5–14, 2018.

[29] I. D. Raji and J. Buolamwini, "Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 2019, pp. 429–435.

[30] R. Xin, J. Wang, P. Chen, and Z. Zhao, "Trustworthy ai-based performance diagnosis systems for cloud applications: A review," *ACM Computing Surveys*, vol. 57, no. 5, pp. 1–37, 2025.

[31] "Ai and society: Challenges and opportunities," The Royal Society, London, Tech. Rep., 2018. [Online]. Available: https://royalsociety.org/topics-policy/projects/ai-and-society/
[32] Stanford HAI, "Artificial intelligence index report 2023," Stanford University, Tech. Rep., 2023. [Online]. Available: https://aiindex.stanford.edu/report/
[33] M. M. Ferdaus, M. Abdelguerfi, E. Ioup, K. N. Niles, K. Pathak, and S. Sloan, "Towards trustworthy ai: A review of ethical and robust large language models," *arXiv preprint arXiv:2407.13934*, 2024.
[34] A. Herrera-Poyatos, J. Del Ser, M. L. de Prado, F.-Y. Wang, E. Herrera-Viedma, and F. Herrera, "Responsible artificial intelligence systems: A roadmap to society's trust through trustworthy ai, auditability, accountability, and governance," *arXiv preprint arXiv:2503.04739*, 2025.
[35] M. Taddeo and L. Floridi, "How ai can be a force for good," *Science*, vol. 361, no. 6404, pp. 751–752, 2018.
[36] D. Li, S. Liu, B. Wang, C. Yu, P. Zheng, and W. Li, "Trustworthy ai for human-centric smart manufacturing: A survey," *Journal of Manufacturing Systems*, vol. 78, pp. 308–327, 2025.
[37] "Ai governance atlas: An overview of global ai governance ecosystems," The Future Society, Tech. Rep., 2022. [Online]. Available: https://thefuturesociety.org/ai-governance-atlas/
[38] A. Søgaard, "Can machines be trustworthy?" *AI and Ethics*, vol. 5, no. 1, pp. 313–321, 2025.
[39] "Blueprint for an ai bill of rights: Making automated systems work for the american people," The White House Office of Science and Technology Policy (OSTP), Tech. Rep., 2022. [Online]. Available: https://www.whitehouse.gov/ostp/ai-bill-of-rights/
[40] P. J. Embí, D. C. Rhew, E. D. Peterson, and M. J. Pencina, "Launching the trustworthy and responsible ai network (train): a consortium to facilitate safe and effective ai adoption," *JAMA*, vol. 333, no. 17, pp. 1481–1482, 2025.
[41] Y. Chinthapatla, "Safeguarding the future: Nurturing safe, secure, and trustworthy artificial intelligence ecosystems and the role of legal frameworks," *International Journal of Scientific Research in Science Engineering and Technology*, 2024.
[42] N. Schlicker, K. Baum, A. Uhde, S. Sterz, M. C. Hirsch, and M. Langer, "How do we assess the trustworthiness of ai? introducing the trustworthiness assessment model (tram)," *Computers in Human Behavior*, vol. 170, p. 108671, 2025.
[43] "Recommendation on the ethics of artificial intelligence," United Nations Educational, Scientific and Cultural Organization (UNESCO), Tech. Rep., 2021. [Online]. Available: https://unesdoc.unesco.org/ark:/48223/pf0000380455
[44] A. Fedele, C. Punzi, S. Tramacere *et al.*, "The altai checklist as a tool to assess ethical and legal implications for a trustworthy ai development in education," *Computer Law & Security Review*, vol. 53, p. 105986, 2024.
[45] G. Stettinger, P. Weissensteiner, and S. Khastgir, "Trustworthiness assurance assessment for high-risk ai-based systems," *IEEE Access*, vol. 12, pp. 22 718–22 745, 2024.
[46] A. Balayn, M. Yurrita, F. Rancourt, F. Casati, and U. Gadiraju, "Unpacking trust dynamics in the llm supply chain: An empirical exploration to foster trustworthy llm production & use," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025, pp. 1–20.
[47] W. Wei and L. Liu, "Trustworthy distributed ai systems: Robustness, privacy, and governance," *ACM Computing Surveys*, vol. 57, no. 6, pp. 1–42, 2025.
[48] Z. Atf and P. R. Lewis, "Is trust correlated with explainability in ai? a meta-analysis," *IEEE Transactions on Technology and Society*, 2025.
[49] A. Balakrishnan, "Leveraging artificial intelligence for enhancing regulatory compliance in the financial sector," *International Journal of Computer Trends and Technology*, 2024.
[50] Y. Nie, S. He, Y. Bie, Y. Wang, Z. Chen, S. Yang, and H. Chen, "Conceptclip: Towards trustworthy medical ai via concept-enhanced contrastive langauge-image pre-training," *arXiv e-prints*, pp. arXiv–2501, 2025.
[51] "The age of digital interdependence: Report of the high-level panel on digital cooperation," United Nations, Tech. Rep., 2019. [Online]. Available: https://www.un.org/en/pdfs/DigitalCooperation-report-for-publication.pdf
[52] B. Kovalevskyi, "Ethics and safety in ai fine-tuning," *Journal of Artificial Intelligence general science (JAIGS) ISSN: 3006-4023*, vol. 1, no. 1, pp. 259–267, 2024.
[53] K. de Fine Licht, "Resolving value conflicts in public ai governance: A procedural justice framework," *Government Information Quarterly*, vol. 42, no. 2, p. 102033, 2025.
[54] G. B. Mensah, "Ensuring ai explainability in clinical decision support systems."
[55] Z. Yu, M. Y. I. Idris, P. Wang, and Y. Xia, "Dancetext: Point-driven interactive text and image layer editing using diffusion models," *arXiv preprint arXiv:2504.14108*, 2025.
[56] M. Wörsdörfer, "Mitigating the adverse effects of ai with the european union's artificial intelligence act: Hype or hope?" *Global Business and Organizational Excellence*, vol. 43, no. 3, pp. 106–126, 2024.
[57] I. Chouvarda, S. Colantonio, A. S. Verde, A. Jimenez-Pastor, L. Cerdá-Alberich, Y. Metz, L. Zacharias, S. Nabhani-Gebara, M. Bobowicz, G. Tsakou *et al.*, "Differences in technical and clinical perspectives on ai validation in cancer imaging: mind the gap!" *European Radiology Experimental*, vol. 9, no. 1, p. 7, 2025.
[58] B. S. Ayinla, O. O. Amoo, A. Atadoga, T. O. Abrahams, F. Osasona, O. A. Farayola *et al.*, "Ethical ai in practice: Balancing technological advancements with human values," *International Journal of Science and Research Archive*, vol. 11, no. 1, pp. 1311–1326, 2024.
[59] K. KN, A. Perrusquia, A. Tsourdos, and D. Ignatyev, "Integrating explainable ai into two-tier ml models for trustworthy aircraft landing gear fault diagnosis," in *AIAA SCITECH 2025 Forum*, 2025, p. 1928.
[60] "Responsible limits on facial recognition technology: Framework for action and case studies," World Economic Forum, Tech. Rep., 2020. [Online]. Available: https://www.weforum.org/reports/responsible-limits-on-facial-recognition-technology
[61] M. Al-kfairy, D. Mustafa, N. Kshetri, M. Insiew, and O. Alfandi, "Ethical challenges and solutions of generative ai: An interdisciplinary perspective," in *Informatics*, vol. 11, no. 3. Multidisciplinary Digital Publishing Institute, 2024, p. 58.
[62] A. Q. Bataineh, A. S. Mushtaha, I. A. Abu-AlSondos, S. H. Aldulaimi, and M. Abdeldayem, "Ethical & legal concerns of artificial intelligence in the healthcare sector," in *2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS)*. IEEE, 2024, pp. 491–495.
[63] B. Schweitzer, "Artificial intelligence (ai) ethics in accounting," *Journal of Accounting, Ethics & Public Policy, JAEPP*, vol. 25, no. 1, pp. 67–67, 2024.

[64] "Global ai action alliance: Accelerating inclusive ai adoption," World Economic Forum, Tech. Rep., 2022. [Online]. Available: https://www.weforum.org/agenda/2022/05/global-ai-action-alliance-inclusive-ai/

[65] Z. Yu, M. Idris, and P. Wang, "Satellitecalculator: A multi-task vision foundation model for quantitative remote sensing inversion," *arXiv preprint arXiv:2504.13442*, 2025.

[66] H. R. Saeidnia, S. G. H. Fotami, B. Lund, and N. Ghiasi, "Ethical considerations in artificial intelligence interventions for mental health and well-being: Ensuring responsible implementation and impact," *Social Sciences*, vol. 13, no. 7, p. 381, 2024.