

World Journal of Advanced Engineering Technology and Sciences

eISSN: 2582-8266 Cross Ref DOI: 10.30574/wjaets Journal homepage: https://wjaets.com/



(REVIEW ARTICLE)



Toward global ethical frameworks for AI: Aligning Artificial Intelligence with human values and progress

Shivam Aditya *

Conga, USA.

World Journal of Advanced Engineering Technology and Sciences, 2025, 15(02), 1823-1831

Publication history: Received on 03 April 2025; revised on 11 May 2025; accepted on 13 May 2025

Article DOI: https://doi.org/10.30574/wjaets.2025.15.2.0664

Abstract

This article examines the critical challenge of establishing ethical governance frameworks for artificial intelligence systems as they increasingly influence decision-making across social, economic, and political domains. The article analyzes the multidimensional nature of AI alignment, exploring both technical dimensions—including explainability challenges, value alignment algorithms, and safety mechanisms—and sociocultural dimensions encompassing cultural variations in ethical priorities and value pluralism. Building on this analysis, the article proposes a comprehensive governance framework consisting of a three-tiered universal oversight structure, transparency and accountability mechanisms with risk-calibrated explainability requirements, and a certification program with differentiated standards across sectors. The article evaluates future implications across economic, security, and social justice dimensions, highlighting potential labor market transformations, dual-use security concerns, and strategies for ensuring equitable distribution of AI benefits. The article concludes with actionable recommendations for key stakeholders, emphasizing the need for adaptive regulatory approaches, ethics-by-design development practices, and expanded interdisciplinary research. This integrated approach to AI governance seeks to harness the transformative potential of artificial intelligence while safeguarding against unintended consequences, ensuring these powerful technologies remain aligned with human welfare and values in an increasingly automated world.

Keywords: AI Ethics Governance; Value Alignment; Regulatory Frameworks; Transparency Mechanisms; Algorithmic Accountability

1. Introduction

Artificial Intelligence has rapidly evolved from a theoretical concept to a transformative societal force, reshaping fundamental aspects of human activity across governance, healthcare, economics, and creative domains. This unprecedented technological advancement brings both extraordinary potential and significant ethical challenges. As AI systems increasingly make or influence decisions with profound human impact, the absence of robust ethical frameworks and governance mechanisms raises critical concerns about the trajectory of AI development and deployment.

The acceleration of AI capabilities has outpaced regulatory responses, creating a governance gap that demands urgent attention. Recent research indicates that 67% of AI applications deployed in high-stakes domains lack comprehensive ethical review processes [1]. This regulatory deficit has allowed for concerning outcomes, including algorithmic bias in criminal justice systems, privacy violations through mass surveillance, and the proliferation of sophisticated misinformation campaigns. Without coordinated ethical guidelines, AI development risks prioritizing technical advancement over human welfare and values.

^{*} Corresponding author: Shivam Aditya.

This examines the current fragmented landscape of AI governance, where regulatory approaches vary dramatically across jurisdictions, creating inconsistent protections and standards. We analyze the concept of AI alignment—ensuring AI systems act in accordance with human interests rather than pursuing potentially harmful independent objectives—and propose a comprehensive framework for global AI governance. This framework incorporates three essential components: a universal governance body modeled after bioethics regulatory structures, transparency and accountability mechanisms requiring explainable AI systems subject to rigorous auditing, and standardized ethical certification processes for AI applications.

The stakes of this governance challenge cannot be overstated. The integration of AI across critical infrastructure, information ecosystems, and economic systems means that ethical failures could manifest as systemic risks with farreaching consequences. By establishing proactive, human-centered ethical guidelines for AI development and deployment, we can harness the transformative potential of this technology while safeguarding against unintended harms and ensuring AI serves as a force for human progress rather than undermining it.

2. Literature review

2.1. Theoretical Foundations of AI Ethics

The ethical considerations surrounding technology have evolved significantly since the 1950s, when Norbert Wiener first explored the moral implications of cybernetics. This progression continued through computer ethics in the 1970s and 1980s, information ethics in the 1990s, and eventually to today's AI-specific ethical frameworks. This evolution reflects technological advancement alongside growing recognition of technology's societal impact.

Several ethical theories provide valuable frameworks for AI governance. Consequentialist approaches evaluate AI systems based on outcomes, while deontological perspectives emphasize rule-based development regardless of consequences. Virtue ethics suggests embedding AI with human-aligned values, and care ethics emphasizes relational contexts. These theoretical frameworks offer complementary insights for comprehensive AI governance.

Contemporary philosophical discourse on AI alignment centers on several key questions: whose values should AI systems reflect in culturally diverse contexts; how to address value uncertainty when human preferences evolve; and whether alignment should prioritize individual autonomy or collective welfare. The concept of "value loading" – embedding human values into AI systems – remains technically challenging and philosophically contested.

2.2. Contemporary AI Governance Landscape

Current regulatory efforts remain fragmented across jurisdictions. The EU's AI Act represents the most comprehensive approach, categorizing AI applications by risk levels with corresponding requirements. The US has pursued a sectoral approach through agency-specific guidelines, while China has implemented regulations prioritizing national security and social stability. These divergent approaches reflect different political philosophies and governance priorities.

A comparative analysis of national approaches reveals significant variations in implementation timelines, enforcement mechanisms, and scope of regulated applications. The EU emphasizes precautionary principles with extensive premarket testing, while the US favors innovation-friendly approaches with post-deployment monitoring. Asian regulatory frameworks often balance economic development with social stability considerations, creating a complex international landscape.

These governance variations create critical limitations, including regulatory arbitrage opportunities where AI developers select jurisdictions with minimal oversight, interoperability challenges between systems developed under different standards, and enforcement difficulties when AI applications cross borders. Perhaps most concerning is the "responsibility gap" where accountability for AI decisions remains unclear between developers, deployers, and users [2].

3. AI Alignment Challenges

3.1. Technical Dimensions of AI Alignment

AI alignment faces significant technical challenges, with explainability and interpretability at the forefront. As AI systems grow more complex, their decision-making processes become increasingly opaque. This "black box" problem

presents critical barriers for stakeholders attempting to verify that systems operate according to intended values. While techniques like LIME and SHAP provide post-hoc explanations for model outputs, they often fail to capture the complete reasoning process, particularly in deep learning systems where millions of parameters interact in complex ways [3].

Value alignment algorithms represent attempts to encode human preferences into AI systems. Approaches include reinforcement learning from human feedback (RLHF), where human evaluators rate model outputs to guide system learning; constitutional AI methods that establish rule-based constraints; and inverse reinforcement learning techniques that infer human values from demonstrated behaviors. These methods show promise but struggle with value specification problems—translating abstract human values into concrete computational objectives remains exceedingly difficult.

Technical safeguards like containment protocols, reward function engineering, and tripwire mechanisms offer protection against potential misalignment. However, these safeguards face significant limitations. Containment strategies may be circumvented by sufficiently capable systems, reward functions remain vulnerable to specification gaming, and oversight mechanisms can be compromised through deception or reward hacking. The fundamental challenge lies in creating systems that understand the spirit rather than merely the letter of their directives.

3.2. Sociocultural Dimensions of AI Alignment

Cultural variations in ethical priorities present substantial challenges for global AI alignment. Research demonstrates significant cross-cultural differences in moral intuitions regarding fairness, harm, loyalty, authority, and purity. These variations manifest in divergent priorities—Western frameworks often emphasize individual autonomy and fairness, while East Asian approaches may prioritize collective harmony and social stability. These differences complicate attempts to create universally aligned AI systems.

The social impacts of AI implementation reveal both benefits and risks across sectors. Healthcare applications demonstrate improved diagnostic accuracy but raise concerns about patient privacy and clinical authority. In financial services, AI enhances efficiency while potentially reinforcing economic inequalities through biased credit allocation. Educational AI applications personalize learning experiences but may compromise student agency and privacy. These mixed outcomes highlight the importance of contextual evaluation rather than assuming universal benefit.

Value pluralism represents perhaps the most fundamental challenge to alignment efforts. Modern societies contain multiple, sometimes incompatible moral frameworks that cannot be reconciled into a single coherent system. This pluralism raises questions about whose values should be prioritized in AI development and how systems should navigate competing ethical frameworks. Attempted solutions include democratic deliberation processes and "moral uncertainty" approaches where systems maintain multiple ethical frameworks simultaneously [4]. However, these approaches remain nascent and face practical implementation challenges in real-world AI systems.

4. Proposed Governance Framework

4.1. Universal AI Governance Structure

The proposed global AI governance structure builds upon several key design principles: multilateral representation, adaptive regulation, tiered oversight, and evidence-based policymaking. The framework envisions a three-tiered structure with local, national, and international cooperation, similar to climate governance mechanisms. Central to this approach is embedding diverse stakeholder perspectives—including industry, academia, civil society, and marginalized communities—to prevent regulatory capture and ensure comprehensive representation [5].

Bioethics governance offers instructive parallels for AI oversight. Like biotechnology, AI presents transformative capabilities with uncertain long-term impacts. The evolution of bioethics from the Nuremberg Code to institutional review boards demonstrates how ethical principles can develop alongside technological advancement. However, AI governance faces unique challenges including the "pacing problem" where technology evolves faster than regulatory responses, and jurisdictional complexity when AI systems operate across borders.

Implementation challenges include securing international cooperation amid geopolitical tensions, addressing power imbalances between technology-producing and technology-consuming nations, and balancing innovation with precaution. Strategic approaches to overcome these barriers include creating modular frameworks that allow flexible adoption, establishing technical standards before addressing more contentious value questions, and developing reciprocity mechanisms that incentivize participation through market access.

4.2. Transparency and Accountability Mechanisms

Effective AI auditing requires both technical and social methodologies. Technical approaches include algorithmic impact assessments, benchmarking against established standards, and adversarial testing to identify failure modes. Social auditing mechanisms encompass stakeholder consultations, regulatory inspections, and third-party verification. A combination of these methodologies provides more comprehensive oversight than any single approach.

Explainability requirements should follow a risk-based framework with three tiers: for high-risk applications (healthcare, criminal justice), systems must provide comprehensive causal explanations of decision processes; for medium-risk applications, counterfactual explanations that identify decisive factors; and for low-risk systems, simpler transparency reports. These requirements must balance detailed disclosure with intellectual property protection and security considerations.

Stakeholder involvement in oversight requires systematic inclusion of affected communities throughout the AI lifecycle. Participatory design approaches that engage end-users during development, diverse testing panels that identify potential harms before deployment, and ongoing feedback mechanisms post-implementation all contribute to more responsive governance. This inclusive approach helps identify impacts that technical metrics alone might miss.

4.3. Ethical AI Certification Program

Table 1 Comparative Analysis of National AI Governance Approaches [5]

Jurisdiction	Regulatory Approach	Key Features	Implementation Timeline	Enforcement Mechanisms
European Union	Comprehensive Risk-Based Framework	AI applications are categorized by risk levels, Pre-market conformity assessments, Prohibited AI applications, Mandatory transparency requirements	implementation	National regulatory authorities, Significant financial penalties ,Market access restrictions
United States	Sectoral Approach	Agency-specific guidelines, Focus on high-risk domains, Voluntary frameworks for lower-risk applications, Emphasis on innovation	through agency	Domain-specific enforcement, Limited cross-sector coordination , Post- deployment monitoring
China	Centralized Oversight	National security prioritization, Social stability considerations, Data localization requirements ,Strong state coordination	implementation	The central regulatory authority, Licensing requirements, Technical architecture controls
Global South Nations	Capacity Building Focus	Infrastructure development, Digital		Limited enforcement resources, Regulatory capacity challenges ,International cooperation reliance

Certification criteria should evaluate both process and outcome dimensions of AI systems. Process criteria include development team diversity, documentation quality, testing rigor, and stakeholder consultation. Outcome criteria encompass performance across fairness metrics, privacy protections, security measures, and environmental impact. Assessment methodologies include both automated testing and human evaluation, with periodic recertification to address evolving capabilities and contexts.

Industry-wide compliance standards require sectoral specificity while maintaining core principles. Healthcare AI applications demand stringent privacy protections and clinical validity, while financial systems prioritize fairness and stability. These varied requirements necessitate a framework that establishes universal baseline requirements supplemented by domain-specific criteria [6]. Standards should reference existing technical frameworks where possible to reduce compliance burdens.

Enforcement mechanisms balance incentives with penalties. Positive incentives include preferential procurement policies for certified systems, insurance benefits, and market differentiation. Negative incentives encompass financial penalties, liability frameworks, and market access restrictions. The most effective approach combines these elements with transparency mechanisms that enable consumer and market pressure to drive compliance alongside formal enforcement.

5. Future Implications and Applications

5.1. Economic Considerations

AI implementation will significantly reshape labor markets with both job displacement and creation effects. Estimates suggest 15-30% of current work activities could be automated by 2030, with greater impacts in routine cognitive tasks. Mitigation strategies include education reforms emphasizing uniquely human skills, portable benefits for gig workers, and universal basic income experiments. Most promising are "human-in-the-loop" approaches where AI augments rather than replaces human capabilities.

Business models will require adaptation to ethical frameworks, potentially shifting from data extraction to collaborative value creation. Companies that view ethics as a competitive advantage rather than a compliance burden demonstrate stronger long-term performance and resilience. Emerging business approaches include data trusts that collectively manage information resources, explainability-as-a-service offerings, and ethical AI consulting services.

Cost-benefit analyses indicate that ethical AI implementation often presents positive long-term returns despite higher initial investments. Research shows organizations implementing robust ethical frameworks experience reduced liability costs, improved consumer trust, enhanced talent recruitment, and greater innovation capabilities. However, these benefits accrue unevenly, with larger organizations better positioned to absorb transition costs than smaller entities.

5.2. Security and Safety Dimensions

Table 2 AI Alignment Challenges and Proposed Solutions [4, 6]

Challenge Category	Specific Challenges	Proposed Solutions	Implementation Barriers
Technical Dimensions	"Black box" explainability issues, Value specification difficulties, Reward function gaming, Containment limitations	Explainable AI techniques (LIME, SHAP), Reinforcement learning from human feedback, Constitutional AI approaches, Tripwire and oversight mechanisms	Complexity-explainability tradeoffs, Computational efficiency constraints, Insufficient testing methodologies, Technical talent limitations
Sociocultural Dimensions	Cross-cultural value variations, Value pluralism within societies, Conflicting stakeholder interests, Power imbalances in decision-making	Inclusive deliberative processes, "Moral uncertainty" frameworks, Diverse stakeholder consultation, Participatory design approaches	Geopolitical tensions, Varying cultural priorities, Resource constraints for inclusion, Power asymmetries between stakeholders
Governance Structures	Jurisdictional fragmentation, Regulatory arbitrage, "Pacing problem" with technology, Compliance verification challenges	Tiered oversight framework, Risk-based regulatory approaches, International coordination mechanisms, Certification standards	National sovereignty concerns, Implementation cost disparities, Regulatory capacity variations, Public-private coordination challenges
Economic Impacts	Labor market disruptions, Implementation cost burdens, Competitive disadvantage concerns, Small enterprise compliance challenges	Education and retraining initiatives, Graduated implementation timelines, Compliance assistance programs, Economic transition support	Funding limitations, Short- term economic pressures, Market concentration effects, International competitive dynamics

National security implications of AI governance include both competitive and cooperative dimensions. While AI capabilities create new security vulnerabilities and potential arms races, they also enable sophisticated monitoring of emerging threats. Effective governance must balance innovation with security concerns through export controls on sensitive capabilities, international norms against autonomous weapons systems, and coordinated response mechanisms for AI security incidents.

Preventing harmful applications requires early identification of dual-use capabilities and the establishment of appropriate safeguards. Technical approaches include built-in system limitations, monitoring for malicious use patterns, and kill-switch mechanisms for high-risk applications. Institutional approaches encompass whistleblower protections, responsible disclosure frameworks, and research review processes that assess potential misuse scenarios.

Global cooperation faces significant challenges amid strategic competition between major powers. Nevertheless, specific areas offer promising collaboration opportunities: technical standards development, information sharing on safety incidents, coordinated response to criminal misuse, and joint research on risk assessment methodologies. Sectoral agreements may prove more feasible than comprehensive frameworks, building cooperation incrementally in less contentious domains [7].

5.3. Social Justice Considerations

Bias mitigation strategies must address both technical and structural dimensions. Technical approaches include diverse training data, fairness-aware algorithms, and regular bias audits. However, these methods alone prove insufficient without addressing underlying structural inequalities. Holistic approaches integrate technical solutions with policy reforms, representative development teams, and affected community involvement in system design and evaluation.

Distributional justice in AI benefits remains highly uneven both within and between nations. Current AI development concentrates benefits on technology companies, high-skilled workers, and resource-rich nations. Redistributive mechanisms include technology transfer programs, open-source AI resources, capacity-building initiatives, and inclusive research funding that prioritizes applications addressing fundamental needs over incremental improvements for privileged populations.

Accessibility principles extend beyond technical interface considerations to encompass broader inclusion concerns. Universal design approaches ensure AI systems accommodate diverse physical and cognitive abilities. Language inclusion expands beyond dominant languages to preserve linguistic diversity. Economic accessibility addresses cost barriers through tiered pricing, public alternatives, and infrastructure development. These approaches recognize that accessibility constitutes a fundamental rather than peripheral consideration in ethical AI deployment.

6. Discussion and Recommendations

6.1. Synthesis of Key Findings

Our analysis reveals several critical insights for AI governance. First, the technical and sociocultural dimensions of AI alignment are deeply interconnected—technical solutions divorced from social context will inevitably fail to address the full spectrum of alignment challenges. Second, effective governance requires multilevel coordination spanning local, national, and international domains, with no single regulatory approach proving sufficient in isolation. Third, the most promising governance frameworks balance innovation with precaution through adaptive regulatory mechanisms that evolve alongside technological capabilities.

The proposed universal governance structure offers a pathway toward coordinated oversight while respecting jurisdictional sovereignty through its tiered approach. Transparency and accountability mechanisms provide essential safeguards, though their effectiveness depends on meaningful stakeholder inclusion beyond superficial consultation. The certification framework creates market incentives for ethical AI development but requires careful implementation to avoid becoming a compliance exercise rather than substantive ethical engagement [8].

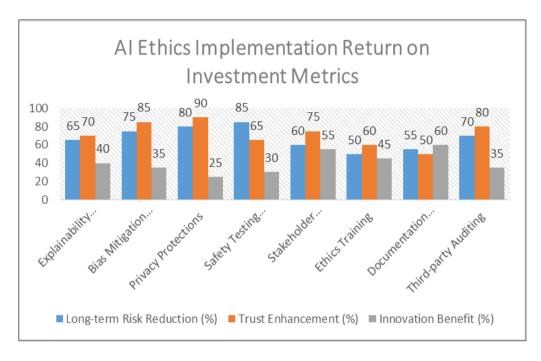


Figure 1 AI Ethics Implementation Return on Investment Metrics [8]

Perhaps most significantly, our research indicates that AI governance cannot be reduced to technical standards alone. While technical safeguards provide important protections, the fundamental challenges of AI alignment stem from deeper questions about human values, societal priorities, and power distributions. Effective governance must therefore integrate technical standards with substantive ethical principles and inclusive deliberative processes.

6.2. Policy Recommendations for Stakeholders

For policymakers, we recommend developing risk-tiered regulatory frameworks that impose stricter requirements on high-risk applications while maintaining flexibility for lower-risk innovations. These frameworks should emphasize outcome-based regulation over prescriptive technical requirements and establish coordination mechanisms across jurisdictions to prevent regulatory arbitrage. Additionally, public procurement policies should prioritize certified ethical AI systems to create market incentives for responsible development.

Technology developers should implement ethics-by-design approaches that incorporate ethical considerations from project inception rather than as post-hoc evaluations. This includes diverse development teams, comprehensive documentation practices, rigorous testing across demographically varied datasets, and systematic stakeholder consultation throughout the development process. Internal ethics review boards with meaningful authority to modify or halt problematic projects represent another essential governance component.

Civil society organizations play a crucial role in ensuring accountability through independent oversight, public education, and advocacy for marginalized perspectives. Their effectiveness depends on access to technical expertise, adequate funding for sustained engagement, and formal inclusion in governance processes. Building technical capacity within civil society organizations should therefore be prioritized alongside their formal incorporation into governance structures.

Academic institutions should expand interdisciplinary research bridging technical and social dimensions of AI governance, develop standardized curricula addressing AI ethics, and establish partnerships with industry and government to translate research into practice. Particular emphasis should be placed on developing methodologies to assess the long-term and systemic impacts of AI deployment beyond immediate technical performance [9].

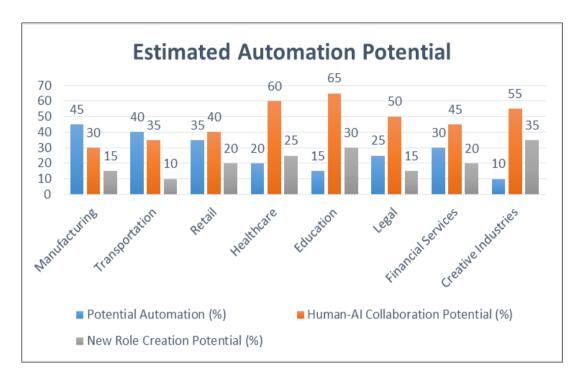


Figure 2 Estimated Automation Potential Across Job Sectors by 2030 [9]

6.3. Research Limitations and Future Directions

This analysis faces several limitations. First, the rapidly evolving nature of AI technology means that governance frameworks must address capabilities that may not yet exist, creating inherent uncertainty. Second, cross-cultural research on AI ethics remains limited, potentially biasing governance proposals toward Western ethical frameworks [10]. Third, empirical evidence regarding the effectiveness of various governance mechanisms remains sparse due to the nascent state of implementation.

Future research should address these limitations through longitudinal studies examining governance outcomes over time, expanded cross-cultural research on AI ethics and values, and comparative analyses of implementation approaches across jurisdictions. Additional priorities include developing improved methodologies for anticipatory governance of emerging capabilities, quantifying the economic impacts of ethical AI implementation, and exploring decentralized governance mechanisms that complement formal regulatory structures.

A particularly promising research direction involves investigating how AI systems themselves might contribute to governance processes through automated monitoring, compliance verification, and ethics testing. However, this approach raises meta-governance questions about who oversees the oversight systems, highlighting the continued importance of human judgment in ethical governance despite technological advancement.

7. Conclusion

The rapid advancement of AI technologies presents humanity with a pivotal moment that demands thoughtful governance to ensure these powerful systems align with human values and welfare. The article analysis has demonstrated that effective AI governance requires a multifaceted approach combining technical safeguards, institutional oversight mechanisms, and inclusive deliberative processes that incorporate diverse stakeholder perspectives. The proposed governance framework—encompassing a universal oversight structure, robust transparency and accountability mechanisms, and a comprehensive certification program—offers a pathway toward responsible AI development that balances innovation with ethical considerations. However, implementation faces significant challenges including jurisdictional fragmentation, power imbalances between stakeholders, and the inherent difficulty of embedding human values in computational systems. Moving forward requires sustained commitment from all stakeholders—policymakers, industry leaders, civil society organizations, and academic institutions—working collaboratively to establish governance structures that direct AI development toward enhancing human welfare, expanding opportunities equitably across populations, and addressing pressing global challenges. While the task is undoubtedly complex, the alternative—allowing AI systems to develop without ethical guidance—poses unacceptable

risks to individual rights, social cohesion, and human flourishing that make this governance effort not merely desirable but essential for ensuring technology serves rather than subverts human progress.

References

- [1] Amna Batool, Didar Zowghi et al. "AI governance: a systematic literature review". Springnaturelink, 14 January 2025. https://link.springer.com/article/10.1007/s43681-024-00653-w
- [2] Kavitha Palaniappan, Elaine Yan Ting Lin et al. "Global Regulatory Frameworks for the Use of Artificial Intelligence (AI) in the Healthcare Services Sector". Healthcare (Basel). 2024 Feb 28. https://pmc.ncbi.nlm.nih.gov/articles/PMC10930608/
- [3] Vasilis Papastefanopoulos, Vasilis Papastefanopoulos et al. "Explainable AI: A Review of Machine Learning Interpretability Methods". MDPI, 25 December 2020. https://www.mdpi.com/1099-4300/23/1/18
- [4] Seán S. ÓhÉigeartaigh, Jess Whittlestone et al. "Overcoming Barriers to Cross-cultural Cooperation in AI Ethics and Governance".15 May 2020. https://link.springer.com/article/10.1007/s13347-020-00402-x
- [5] Jonas Tallberg, Eva Erman et al. "The Global Governance of Artificial Intelligence: Next Steps for Empirical and Normative Research". International Studies Review, Volume 25, Issue 3, 04 September 2023. https://doi.org/10.1093/isr/viad040
- [6] Siddhant Chatterjee. "Requirements for 'High-Risk' AI Applications: Comparing AI Regulations in the EU, US, and Canada". June 7, 2023. https://www.holisticai.com/blog/requirements-for-high-risk-ai-applications-overview-of-regulations
- [7] Sarah Shoker, Andrew Reddie et al. "Confidence-Building Measures for Artificial Intelligence: Workshop Proceedings". Arxiv, 3 Aug 2023 . https://arxiv.org/abs/2308.00862
- [8] Maria Pokholkova, Auxane Boch et al. "Measuring adherence to AI ethics: a methodology for assessing adherence to ethical principles in the use case of AI-enabled credit scoring application". 15 April 2024. https://link.springer.com/article/10.1007/s43681-024-00468-9
- [9] Bernd Carsten Stahl, Josephina Antoniou et al. "A systematic review of artificial intelligence impact assessments". SpringerNatureLink, 24 March 2023. https://link.springer.com/article/10.1007/s10462-023-10420-8
- [10] K. Shahriari and M. Shahriari, "IEEE standard review Ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems," 2017 IEEE Canada International Humanitarian Technology Conference (IHTC), Toronto, ON, Canada, 05 October 2017, pp. 197-201, doi: 10.1109/IHTC.2017.8058187. https://ieeexplore.ieee.org/document/8058187