JADHUR

Vol.4, Issue.3, (2025)

Journal of Advances in Humanities Research

https://doi.org/10.56868/jadhur.v4i3.310

Artificial Intelligence Literacy and Ethical Digital Governance: Pathways of Multi-Stakeholder Collaboration and Value Alignment

Xu Hao[1], Nurul Liyana Mohd Kamil [1*]

1 Faculty of Business and Economics, Universiti Malaya, 50603 Kuala Lumpur, Malaysia

Article Information

Article Type: Review Article

Dates:

Received: 01 May 2025

Revised: 03 August 2025

Accepted: 09 August 2025

Available online: 22 September 2025

Corresponding Author: Narul Liyana Mohd Kamil

Email: nurulliyana@um.edu.my

ABSTRACT

The need for ethical governance based on responsibility, transparency, and equity has increased due to the rapid integration of artificial intelligence (AI) in business, education, and governance. This study explores the relationship between value alignment, multi-stakeholder engagement, and AI literacy as key pillars of moral digital governance. Based on public value theory and stakeholder theory, the study employs a qualitative interpretive design that combines comparative case analysis and grounded theory. Between 2019 and 2025, data on governance systems in the Global South, China, and the European Union were gathered from academic publications, institutional reports, and international policy instruments. According to research, value alignment serves as the normative outcome that ensures consistency between AI systems and social ethics, collaboration serves as the working tool that transforms literacy into ethical governance, and AI literacy serves as the cognitive foundation for evidence-based engagement. While fragmented or top-down models of governance fell short in terms of accountability and inclusivity, regions such as the EU and China that integrated AI literacy and participatory governance into their institutional framework demonstrated greater ethical coherence and public trust. The study offers a strong theoretical and applied framework that demonstrates how literacy-based collaboration results in moral and value-based governance. The study recommended trust-based and participatory discussion instead of regulatory compliance to AI governance. In addition the study recommended that the stakeholders such as educators, policymakers, and technology develop the models that connect the most important components that are, literacy, cooperation, and value alignment to develop human-centered and socially acceptable AI technologies in the emerging digital era.

Keywords: Artificial Intelligence Literacy, Ethical Digital Governance, Multi Stakeholder Collaboration, Value Alignment, Stakeholder Theory, Public Value Theory, Responsible AI

1. INTRODUCTION

At an advanced stage of development, Artificial Intelligence (AI) is influencing every aspect of life, including global information sharing, business, international relations, corporations, social structure, and governance. The adoption of AI technologies, such as Deepseek, ChatGPT, and Sora, significantly influences the generation of information across health, education, news, and justice (Hussein et al., 2025). On the one hand, AI is helping generate information, while on the other hand, these technologies have raised serious ethical issues. Some of the major challenges posed by these advanced technologies include information distortion, algorithmic bias, and data misuse (Hidayat and Muis, 2025). AI literacy is not just the ability to master AI technology, but also the ability of citizens to participate in governance, criticize algorithmic power, and provide ethical feedback in a digital society. It is a cognitive, critical, and practical "socially embedded literacy" system (Zawacki-Richter et al., 2019; Chu & Dong, 2024).

In addition, with the widespread application of AI systems in highly sensitive social scenarios, the contradiction between AI systems and human value systems has become increasingly intense, which has aroused great attention from academic and policy circles to the issue of "value alignment" (Russell, S., & Norvig, 2021 . Whereas existing studies have focused on AI literacy at the individual or organizational level (Mills et al., 2024), fewer have investigated how it intersects with multi-stakeholder collaboration, especially in reconciling disparate values across sectors (Güngör, 2020; Prikshat et al., 2022). Current literature either addresses technical capabilities in AI literacy or mechanisms for governance of collaboration (MacDonald et al., 2019), but does so infrequently by interweaving the two views. By defining the gap, the current research highlighted the necessity of a more integrated framework that captures the interaction between AI literacy and collaboration processes to ensure value alignment.

It is challenging for a single leading figure to create an integrated policy when the governance process lacks transparency, which limits understanding, accountability, and trust in decision-making. In order to overcome these issues, the developers, administrators, and policy makers are focusing on coordination among various stakeholders to develop such systems and frameworks that minimize these ethical concerns and produce transparency and accountability (Irawati et al., 2023; Birdayanthi et al., 2025). This collaboration will not only support the distribution of resources and responsibilities but also promote social integration and ethical standards.

The existing literature has focused on AI literacy at the individual or organization level, fewer have investigated how it meets with multi-stakeholder collaboration, especially in reconciling disparate values across sectors (Jiang et al., 2024; Hingle, A., & Johri, 2025). To fill these gaps the current research highlighted the necessity of a more integrated framework that captures the interaction between AI literacy and collaboration processes to ensure value alignment.

1.1 Conceptual definition of the key terms

Artificial intelligence (AI) literacy can be defined as the knowledge, skills, and attitudes that enable people to understand, engage with, and critically assess AI technologies in daily life. It involves understanding the workings of AI systems, their potential advantages and drawbacks, and the technology's social and ethical consequences (Miao & Holmes, 2023). Ethical digital governance entails setting principles and policies, and frameworks that are used to handle digital technologies and data in a

responsible, transparent and fair manner. It focuses on responsibilities, justice, privacy, and inclusivity of both the design, implementation, and management of digital systems (Floridi and Cowls, 2019). A multi-stakeholder approach involves bringing together and involving various actors in decision-making, including governments, private-sector organizations, civil society, academia, and international institutions. This would contribute to the development of shared accountability and joint problem-solving, particularly in global governance and technology (Pauwels, 2020; United Nations, 2022). The value alignment is the mechanism of making artificial intelligence behave in accordance with human values, ethical standards, and social objectives. It entails creating AI algorithms to achieve a human intent in machine goals to avoid inadvertent injury or bias (Jobin et al., 2019).

1.2 Theoretical Framework

The promotion of transparency, fairness, justice and accountability is one of the urgent needs in the advanced technological era. The present study aims to address this need by integrating artificial intelligence (AI) with stakeholder collaboration to promote governance and value alignment. The study is grounded on the two most essential theories, Public Value theory and Stakeholder theory. Together, these theories highlighted the importance of coordination and collaboration in promoting social accountability and improving administrative processes and governance.

According to stakeholder theory, one of the primary responsibilities of governance systems and organizations is to foster moral obligations by promoting an environment grounded in moral standards for all stakeholders engaged in the virtual ecosystem (Donaldson & Preston, 1995). In the domain of AI supremacy, it interprets administration as a conciliatory process involving multiple actors, where legality emerges from the considered contributions of diverse actors (Batool et al., 2025). Attaining moral supremacy cannot depend on individual governance; it must arise from shared dialogue, confidence-building, and collective responsibility across official and social structures (Kraus et al., 2021). AI Literacy plays a key role as a promoter, boosting investors' ability to participate efficiently in these developments. Investors who are well-educated in both the practical and ethical aspects of AI can identify machine learning biases, review administrative systems, and advocate for transparency and accountability (Ng, 2021). While a lower level of AI literacy creates problems of data disorder and integrity. Thus, Digital Competence is vital for bringing investors to a mutual consideration, empowering informed discussion, and linking individual consideration with collective moral action. Stakeholder Theory, therefore, provides the operational facet of the outline explaining who is associated and how these communications govern moral consequences (Bridoux & Stoelhorst, 2022; Pies, 2023).

Public Value Theory (Moore, 2013; Sabatier and Weible, 2014) complements a regulatory dimension of power, highlighting the need to improve social welfare rather than merely focusing on functional productivity. The theory pointed out that legitimacy arises when the governance system produces results in accordance with universal moral standards and principles such as impartiality, inclusiveness, and justice (Bozeman, 2007). In the AI field, the personal liberties, equality, and community confidence that emphasize clarity, comprehensibility, and hands-on support are built on moral virtual supremacy in virtual spheres (Floridi et al., 2018).

This collaboration among academia, industry, developers, organisations, and the general public will encourage negotiation among these stakeholders and stabilise conflicting standards such as competence vs confidentiality or modernization versus the law (Winfield and Jirotka, 2018). These theories pointed out that stakeholder integration is the key to achieving ethical virtual administration. Also, these theories claimed that operational involvement, virtual competencies, information sharing, moral obligations and trust building among stakeholders is possible by teamwork, promoting universal laws and principles, and having public conversations (Novelli, et al., 2023; Jobin et al., 2019).

Eventually, the outline exemplifies moral virtual administration as a cooperative co-production method based on collective ethical vision and supportive systems. By merging the inclusivity focus of Stakeholder Theory with the legitimacy and public good emphasis of Public Value Theory, this research situates AI governance within a framework of collective intelligence, ethical coherence, and institutional trust, all of which are essential for promoting responsible AI in today's digital landscape.

1.3 Research questions

This study focuses on two key factors in the governance of artificial intelligence: "artificial intelligence literacy" and "digital ethical governance." To enhance clarity and applicability, the analysis has been narrowed down to examine the interactions within multi-stakeholder collaborations and the methods for aligning values. With this emphasis, the research explores the following practical questions:

- How are targeted efforts (e.g., in-workplace training or community education programs) to make the public AI literate sufficient for effective participation in governance mechanisms?
- How do different governance actors' policymakers, technology developers, and civil society collaborate to conceive and implement ethics-based digital governance mechanisms in real institutional contexts?
- How are technology systems structured such that value alignment is achieved in practice, particularly in the face of competing priorities in social, economic, and cultural spheres?

1.4 Research purpose and significance

The primary objective of this study is to examine how Artificial Intelligence (AI) literacy contributes to the development of ethical digital governance through the mediating pathways of multi-stakeholder collaboration and value alignment. Specifically, the study aims to (1) conceptualize AI literacy as both a technical and ethical competency essential for responsible participation in governance processes; (2) analyze the role of multi-stakeholder collaboration in translating literacy into inclusive and accountable governance practices; and (3) explore how value alignment serves as the normative mechanism ensuring coherence between technological innovation and societal ethics. By integrating Stakeholder Theory and Public Value Theory, the study seeks to develop a comprehensive framework that explains how informed participation and shared moral reasoning can strengthen the legitimacy of AI governance.

The significance of this study lies in its theoretical and practical contributions to the evolving discourse on responsible AI governance. The study is important because it bridges intellectual considerations with public and organizational integrity to overcome the gap between academic studies and administrative

learning. Moreover, the study recommended knowledge-based participation for policy makers, educators, and online influencers to foster moral and social responsibility among stakeholders and the public engaged in the virtual ecosystem. Hence, this study makes a significant contribution to developing such administrative frameworks that are technically sound and grounded in universal moral standards.

2. LITERATURE REVIEW

On one hand, the growing AI technologies have transformed the mode of information processing, while on the other hand, they have introduced many complex ethical issues. Scholars and policymakers have come to recognize the importance of the public in interpreting, engaging in critical thinking, and making ethical decisions when dealing with AI. Therefore, the concept of artificial intelligence literacy has gained new theoretical and governance aspects. At the same time, the moral unrest of AI systems also triggered the recognition of value alignment as a crucial element of AI ethical management. AI literacy, ethical online governance, and value alignment are interdependent and inseparable. Such constructs are shown to be mutually influential on the standardization, legitimacy, and acceptance of the use of technology in society (Long and Magerko, 2020; Su et al. 2023). AI literacy enhances an individual's ability to comprehend and interact critically with intelligent systems, and ethical digital governance helps ensure that such technologies are developed and governed in accordance with the principles of accountability, fairness, and transparency (Tahaei et al., 2023). Moreover, value alignment is considered the ethical compass of the human-AI relationship, which dictates how societies can govern technological innovation and respond to the challenges posed by new organizations and ethical questions. Together, these dimensions will decide how future societies will strike a balance between innovation and ethics in the digital era (Kim et al., 2021).

Therefore, this section will conduct a systematic literature review and comprehensive analysis on the connotation reconstruction of AI literacy, the types and causes of digital ethical risks, global governance models and value alignment theory. By doing this, the research aimed to establish the theoretical basis, foundation, and problem context for the current research.

2.1 Connotation reconstruction and governance significance of artificial intelligence literacy

Artificial intelligence literacy is a compound capability structure that encompasses a basic understanding of AI principles, ethical judgments about AI risks, critical thinking about platform behavior, and practical ability to interact and collaborate with AI systems (Ng et al., 2021). Long and Magerko (2020) define it as "the set of capabilities that enable individuals to understand, use, critique, and participate in algorithmic systems in an AI-dominated society." scholars like Zou and Schiebinger (2018) and Crawford and Calo (2016) posit that it should further involve ethical justification, critical consciousness, and socio-political comprehension, tracing academic debates regarding its scope and emphasis.

Zawacki-Richter et al. (2019) indicate that AI literacy includes four primary components: technical knowledge, data literacy, ethical literacy, and social engagement. These ideas are strongly supported, but there is still debate about how they can be applied in other educational systems. As a case in point, Global North nations, including the UK and Singapore, have incorporated AI literacy into their education systems and emphasize systematic inclusion and exposure at a young age. Instead, Iran, India, and the Global South are interested in community-based AI training to address digital inequities (Marzdar, 2025). This comparison suggests the various geopolitical and socio-economic factors that affect the various strategies, meaning that AI literacy is founded on the status of development and culture rather than international

standards. As an individual, the AI illiteracy could often leave the user with no insight into how the system of algorithms operates, the harm of such technology, and any other prejudices that are prone to the technology. This lack of knowledge could lead to an accidental increase in technical risks (Wu et al., 2025). As the analysis of generative artificial intelligence literacy indicates, AI literacy must not merely comprise a basic understanding of algorithms and systems, but also encompass the understanding of misinformation detection processes, ethical value judgment, and risk tracking on platforms and in action (Liu et al., 2025). This is a full set of skills necessary for society's involvement in the decision-making process for artificial intelligence. AI literacy in China can be considered one of the most prominent qualities of humanistic embeddedness, integrating technological knowledge, ethical implications, and social responsibility. As the influence of generative artificial intelligence on education, justice, governance, and culture continues to grow, it is no longer solely an educational concern but also a significant aspect of social governance. This transformation reveals the importance of ethics and institutional framework in aiding effective decision-making, collaboration amongst stakeholders and congruence between the technological development and social ideals. China's strategy may serve as an excellent example of how AI can be effectively implemented in society (Zhu, 2024). AI literacy serves as an agent of ethical digital governance and value-based coordination among actors, directly aligning with the overall aims of this paper.

Risk prevention mechanism: AI awareness enables individuals to be conscious and decompose the ethical risk of algorithmic bias, deepfakes, and data breaches. According to Alharbi. (2021), this knowledge infuses responsible, informed awareness that must be in place to address these threats and prevent mess in online communication systems.

Participatory co-governance capacity: Participating in digital ethics conversations, expressing concerns, and engaging in governance processes are ways for citizens to develop a thoughtful understanding of AI and articulate their needs. Sadat (2025) adds that such capability is what turns the public not into passive technology consumers, but a source of active engagement in the creation of ethical norms and thus the foundation of cooperation between different stakeholders in the field of digital governance

Practical background on value alignment: AI literacy can help those who use AI establish a collective understanding of what is and is not appropriate and ethical in its use by fostering continuous dialogue and contemplation. Karimov and Saarela (2025) argue that this common moral thinking establishes social consensus, offering a feasible basis for value congruence between AI systems and societal governance objectives. In the global tide of "responsible AI" governance, artificial intelligence literacy is more than an individual capability; it encompasses the entwined capabilities of nations, platforms, and governance systems. Individual literacy directs responsible use and ethical awareness, whereas national and platform capacities institutionalize these values through policy and design. China's New Generation AI Development Plan, for example, demonstrates how concerted policy frameworks integrate personal, institutional, and systemic literacy to promote ethical and responsible AI governance (State Council Information Office of the People’s Republic of China, 2025).

2.2 Analysis of the types and mechanisms of digital ethical risks

While generative artificial intelligence breaks the boundaries of traditional technologies, it also brings unprecedented ethical risks. These risks are not only reflected in privacy violations and misinformation at the individual level, but also point to the erosion of governance mechanisms, the erosion of value bases, and the unequal reproduction of social structures. These risks stem from the compound effect of embedding

platform logic, social structure, and governance gaps rather than the "out of control" nature of the technology itself (Beer, 2017; Selbst et al., 2019). This section examines the main ethical risk categories and formation mechanisms of AI systems from three angles. (1) the technological dimension which deals with algorithmic bias, data quality and transparency of the system; (2) the human or social dimension, which deals with how people use it, whether they are dependent on a system or not and the impact the system has on society; and (3) the institutional or governance dimension, which deals with regulatory framework, accountability and ethical supervision.

The breakthrough in generative AI's information-generating ability makes it not only a tool for content production but also a "weaver" of information cognition. Large model systems such as ChatGPT and Sora can generate multiple rounds of complex sentences, images, audio & video, but their "language simulation" does not imply "fact truth" (Bommasani et al., 2022; Ji et al., 2023). In the absence of reliable data and with biased training datasets, large models often generate hallucinatory content that is "plausible but wrong" (generative AI systems like big language models generate output that is superficially coherent, fluent, and factually plausible, but incorrect, made-up, or not based on actual data) (Islam Tonmoy et al., 2024).

According to Rawte et al.(2023) lack of AI literacy can amplify the "technological hallucination" (The phrase technological hallucination is used to describe a phenomenon wherein artificial intelligence systems, large language models (LLMs) and generative AI, generate as output text or other such things that are grammatically correct and sound plausible but factually in error or wholly made up) effect, leading users to place excessive trust in AI systems, mistaking them for neutrality and objectivity, which in turn weakens their ability to recognize platform manipulation and commercial manipulation (Romanishyn, 2025). AI literacy serves as a link between public decision-making in democracies and ethical consensus, as well as the capacity to comprehend technology. Hristovska (2023) has noted that the issue of information disorder surrounding the dissemination of AI-generated content has not been adequately addressed by the platform governance mechanisms currently in operation.

On websites like YouTube and TikTok, for instance, recommendation algorithms customize AI-generated content to users' emotional and behavioral preferences. Such algorithmic personalization, sometimes referred to as the "emotion algorithm," intensifies information cocoons and echo chambers where users are continuously exposed to agreeable content. As a result, this weakens critical thinking and fosters a "cognitive distortion field" that spreads prejudice and false information online.

Furthermore, the issue of misinformation is not merely a "corpus problem (biased data)"; rather, it is a manifestation of the intricate game between platform responsibility, user capabilities, and algorithm induction since AI systems are used in high-risk contexts like education, government affairs, and healthcare (Saeidnia et al., 2025).

In order to improve information governance capabilities, it is necessary to implement a systematic "knowledge transparency" (openness of AI systems with regard to revealing how information is being produced, namely data sources, algorithms, and reasoning processes) and "semantic verification" (checking AI outputs to know about the accuracy and actual consistency) mechanism based on the improvement of AI literacy. According to Rajkomar et al. (2019), extensive clinical decision support systems such as Google's DeepMind Health demonstrated that their lack of semantic verification, that is, confirming that AI-generated data is in tune with actual-world meaning and context, led to patient information being misinterpreted even when AI models had high predictive accuracy. This highlighted how crucial it is for

users to be AI-literate and have knowledge transparency (a clear understanding of how AI produces outputs) to comprehend and properly regulate such systems.

High-frequency data collection, analysis, and modelling are essential components of AI systems, and the underlying logic of these systems is defined by "implicit intrusion" (Klenk, 2023). The platform trains the model using user clicks, browses, and inputs, and its algorithmic behaviour occurs in the absence of clear consent procedures and legal authorization. For example, large models construct "digital portraits" by continuously tracking and simulating user behavior, which, in turn, affects their decision-making, thereby creating a form of hidden manipulation. As Milano et al. (2020) point out, AI-driven personalized recommender systems tend to shape user behavior rather than serve user preferences, given information asymmetry and algorithmic inexplicability.

Prajescu and Confalonieri (2025) noted that the introduction of AI-assisted trial technology into the judicial system has highlighted the importance of data collection, model training, legitimacy, and interpretability. Technical intervention may alter adjudication logic, erode judges' subjectivity, and create the illusion of "quasi-personality" and "quasi-judgment." This describes how computer programs can appear to imitate the way people think and judge right from wrong. This creates the mistaken belief that the machines possess human-like decision-making power or a personal capacity for judicial judgment (Baum, 2020). In highly sensitive fields such as health care, education, and finance, the absence of data governance mechanisms may pose serious ethical and security risks. To address ethical and security issues, China employed an AI-aided sentencing system that uses historical case data to produce penalty recommendations. Although these systems make the courts more efficient and consistent, research indicates that judges may over-rely on algorithmic outputs, thereby undermining their capacity to reason and make ethical decisions independently (Socol et al., 2024). This example illustrates the general concern that without strong data governance and ethical controls, AI systems have the potential to transform professional judgment, increasing risks in other high-stakes domains like healthcare and finance.

However, due to the triple identity of platform companies in the AI ecosystem as technology developers, service operators and data controllers, the boundaries of their responsibilities are blurred, and the pressure of governance is shifting, resulting in the weakening of public supervision capabilities. This phenomenon of "platform responsibility hollowing out" has become one of the key obstacles to AI ethical risk management. This dilution of accountability is a significant hurdle for effective AI ethical risk management, as it undermines accountability within the AI governance framework (Novelli, 2023).

In social-scenario applications, AI is not a neutral technical tool but a "value container" for social construction.

The “value container” term emphasizes that artificial intelligence-based computer systems are not merely impartial tools; instead, they are designed to replicate and incorporate existing moral, social, and cultural principles. It has been based on the Social Construction of Technology (SCOT) perspective, which argues that both the institutional environment, social norms, and human decisions ultimately determine how any technology is created and used (Pinch and Bijker, 1984). All training data, model architecture, and usage contexts are historically, culturally, and structurally grounded; such structures are often not visible to the general population (Foka et al., 2025). As numerous studies have demonstrated, AI systems are racially, gender, and class-biased in scenarios such as image recognition, credit approval, and hiring. Such biases are due to either an opaque model design or an Imbalanced training sample (Obermeyer et al., 2019;

Mehrabi et al., 2021). AI literacy has become an essential competency in the digital era because it enables individuals to understand, assess, and practice with AI technologies responsibly. This is a necessary skill because AI is increasingly influencing governance, employment, and education. Moreover, as the digital divide is driven by low levels of digital access and uneven technological exposure, it is essential to educate individuals in AI literacy to bridge the divide and ensure the inclusion of people in AI-driven economies in developing countries (UNESCO, 2023).

The model of AI literacy in developing countries proposed by Kathala et al. (2025) highlights the significant roles of resource scarcity, cultural peculiarities, and differences in educational systems in promoting AI literacy. Beyond providing a systematic approach to developing local educational programs and policy modifications, their study suggests that, to advance AI education, policymakers should integrate social justice and technical skills.

2.3 Comparing global models and creating multi-agent collaborative frameworks

As AI is a global technology, the issues related to its governance and ethical integrity are global in scope. Priorities related to these concerns vary. For instance, according to the OECD Principles and the European AI Act, the key issues associated with AI technology are accountability and transparency, whereas in Asian and Global South approaches, inclusivity and limited understanding and knowledge are the primary concerns (UNESCO, 2024).

To ensure a human-centered approach, the Organisation for Economic Co-operation and Development AI Principles (2019) set key ethical standards for global governance. These moral standards are based on five pillars: inclusive growth, human-centered values and fairness, transparency, clarity, accuracy, and accountability. The United Nations and G20 followed these five principles as cross-cutting themes throughout their initiatives (OECD, 2019; Veale et al., 2021).

Similarly, the European Union Artificial Intelligence Act (2022) legally operationalizes ethics by requiring compliance with these ethical standards. The EU AI Act categorizes AI systems into two major categories: one based on risk, and the other focused on mechanisms to promote transparency obligations and human oversight requirements (Consilium, 2024; Ebers, 2025).

The United States (US) promotes innovation through an Open-standard policy. Hence, there are variations in policies and principles related to AI technology, as highlighted by the discussion above. As the US promotes a fragmented, open-standard approach, the EU emphasizes the risks involved, and Asian and Global South nations favor a development-oriented model, focusing on AI literacy (Robles & Mallinson, 2023).

By emphasizing the integration of local social norms and capacity building, these strategies diverge from the West's focus on compliance or innovation and ultimately highlight context-sensitive ethics.

2.4 Regional-wise comparative overview of AI governance approaches

Table 1 presents a comparative analysis of the major models of AI governance, evaluated with respect to regulatory structure, policy instruments, operational characteristics, and principles. This multidimensional comparative analysis provides a clear understanding of the impact of various institutional traditions on the regulatory policies for addressing common moral issues such as accountability,

transparency, human rights, and distributive justice. The comparison highlights three main AI governance models (1) the precautionary, law-based paradigm of AI governance in the EU; (2) the market-based, decentralized paradigm of AI governance in the US; and (3) the development-oriented, capacity-building paradigm of AI governance in Asia and the Global South, with its emphasis on inclusivity and digital literacy.

Table 1. Comparative Overview of AI Governance Approaches

Region	Governance Approach	Key Characteristics	Normative Orientation / Regulatory Logic
European Union	Comprehensive regulation (EU AI Act)	Risk-based classification, binding legal obligations, conformity assessment, strong oversight	Precautionary, compliance-driven, rights-protective governance
United States	Fragmented policy model	Open standards, innovation-focused, sector-specific regulation, decentralized enforcement	Market-oriented, innovation-driven, flexible regulatory governance
Asia & Global South	Context-sensitive, development-oriented approach	Inclusivity, cultural sensitivity, AI literacy, institutional capacity building	Developmental, distributive, and capacity-building governance

Table 1 illustrates that the European Union institutionalizes a precautionary paradigm of governance, based on the enforcement of legal responsibility and the protection of fundamental rights. The United States, on the contrary, has a more regulatory stance whereby technology competitiveness and adaptability of the market are favored, leading to a decentralized and relatively loose structure. Meanwhile, developmental equity, social preparedness, and situational ethics are anticipated in governance across Asia and the Global South, and both regions prioritize long-term capacity building over short-term regulatory inflexibility.

Such an organized comparison shows that AI governance variance is not just regulatory disparity but also an indication of more normative commitments and institutional path dependencies. The results thus support the conclusion that the multi-agent collaborative governance design can incorporate compliance protection, innovation promotion, and an inclusive growth agenda within a globally integrated yet localized framework.

2.5 Country-Level Comparative Analysis of AI Governance

Table 2 presents a country-by-country comparison of AI governance structures in China, the European Union, and the United States, explicitly including implementation. The integration of governance

strategies, the nature of regulation, and the system's constraints enables a more nuanced evaluation of the merits and flaws of each model using the table.

The analysis identifies three AI governance models that include the multi-regulatory, state-based model in China, the precautionary, legally binding, and risk-based model in the EU, with a focus on rights and oversight, and the decentralized, innovation-oriented model in the U.S., focusing on flexibility and market-oriented guidance.

The added column of Challenges highlights the fact that the effectiveness of governance relies not merely on the design of regulations but also on the extent of coordination between institutions, and on their enforcement capabilities and coherence, bringing the analysis off the normative comparison to the realities of real-world implementation.

Table 2. Country-Level Comparative Analysis of AI Governance

Country/Region	Governance Approach	Key Characteristics	Challenges
China	Multi-regulation	Legal, ethical, and political guidance; emphasis on fairness and social ethics; strong state coordination	Fragmented administrative responsibilities; limited public participation; uneven enforcement
European Union	Risk-based regulation (EU AI Act)	Binding legal obligations; structured risk classification; transparency and human oversight requirements	Complex compliance procedures; regulatory burden on SMEs; slower adaptation to rapidly evolving AI technologies
United States	Innovation-focused, fragmented	Open standards; encouragement of R&D; sector-specific and flexible regulatory guidance	Limited federal coordination; gaps in accountability mechanisms; uneven oversight across sectors

Table 2 highlighted that each country has a different institutional logic and relations between the state and the market. The system has a high degree of central coordination, but implementation inconsistency and lack of transparency through participation, which is evident in China. The EU framework is most formalized and rights-protective, but too complex to provide compliance burdens and to adapt fast to changing technologies. The U.S. model promotes innovation, responsiveness, and leadership in the private sector but suffers from coordination failures and accountability lapses due to regulatory fragmentation.

Both Table 1 and Table 2 provide complementary comparative information. Table 1 reveals general regional governance paradigms and normative orientations, whereas Table 2 discusses national-level regulatory frameworks and problems of regulatory implementation. Together, these tables illustrate how general principles of governance are implemented in concrete institutional practices and note not only similar ethical purposes but also discrepancies in the application of regulations. This combined analysis confirms that there is no unilateral balance involving regulatory certainty, innovation flexibility and participatory legitimacy. This analogy underscores the necessity of a multi-agent collaborative governance model that incorporates legal enforceability, normative diversity, and technological flexibility and involves all stakeholders across jurisdictional levels.

Southeast Asian countries, such as Indonesia and Malaysia, focus on coordinating among stakeholders, including educational institutions, communities, and civil society, to promote AI literacy and ethical awareness. These countries ensure inclusive digitalization by stressing equal access and a participatory approach in AI literacy (Xu et al., 2024).

This participatory orientation complements the development-oriented governance logic identified in Tables 1 and 2, in which inclusivity and capacity building are central regulatory priorities rather than strict compliance mechanisms.

However, India promotes the integration of AI into public services by emphasizing modernization and efficiency, while giving less attention to civic engagement and ethical deliberation (Lakshitha et al., 2025). This is indicative of a techno-developmentalist trajectory where efficiency in administration is seen as more important than ethical systems of governance.

Similarly, countries like Bangladesh and Pakistan are still in the early stages of AI governance; their policies focus on digital infrastructure and economic competitiveness. However, these countries provide less attention to ethical risk management mechanisms (Karmakar, 2024; Muhammad et al., 2025). This is consistent with the global South governance paradigm discussed above in which infrastructural preparedness frequently preludes formalized accountability.

Similarly, Nepal and Sri Lanka have initiated the integration of AI into the educational curriculum and teachers' training on AI, with support from UNESCO. However, these countries lack proper execution of these strategies (UNESCO, 2024). This implementation gap is equivalent to the capacity-execution gap in Table 2, where policy statements are not always realized in practice.

In summary, the literature highlights structural disparities in ethical governance and value alignment across regions. The literature shows that the core areas of focus in South Asian countries are capacity development of organizations and economic modernization, whereas Southeast Asian countries have paid closer attention to inclusive and participatory governance. These differences illustrate how variations in the regional governance logics, compliance-driven, innovation-driven, or development-oriented, are reflected in institutional practice and order of policy.

Wright (2024) said that despite advocating "humanistic AI" and "global responsibility, Japan's AI ethics practice still maintains a conservative structure, with insufficient participation of women and minorities, a lack of implementation system, and ethical principles mostly remaining at the document level. For instance,

gender imbalance undermines the work of groups dedicated to ethical AI. The Humanistic AI Social Principles Group includes only 13.8% women, and all Moonshot project managers are men. This discrepancy highlights a fundamental contradiction between the commitment to diversity and actual governance practices. Such structural exclusion ultimately erodes the inclusiveness and legitimacy of global AI ethics initiatives. This instance also supports the distinction between normative and institutional realizations of commitment and the realization recognized in comparative analyses of governance across regions.

In Iran, AI policy is heavily influenced by national security logic, with a highly centralized regulatory system that lacks space for ethical debate from multiple perspectives (Atwood, 2025). This model reinforces government authority but also undermines social resilience in addressing AI risks.

In contrast to the participatory or market-driven models mentioned above, this centralized approach demonstrates how security-based governance can limit the process of pluralistic value alignment.

2.6 Theoretical dimensions (comprehensive theoretical construction framework) and practical challenges of value alignment

Based on the comparison analysis above of governance, the focus now turns from institutional models to the theoretical foundation of value alignment as the normative basis for the connection between governance structures and AI system behavior.

In the context of artificial intelligence systems increasingly embedded in human society, ensuring that their behavioral goals do not deviate from human values, ethical principles, and true intentions has become one of the most urgent and complex issues in AI ethical governance. "Value Alignment" is one of the shared topics of interest across AI, computer science, and policy governance (Russell et al., 2015). Though “Value Alignment" is a new concept, its basic purpose has historical roots.

Historically, Value Alignment has confirmed that AI functions are related to human values. This concept originated from Isaac Asimov's "Three Laws of Computing” and was further developed through cybernetics and ethical computing frameworks, which examined how human behavior is controlled by machines to achieve standardized objectives (Zwitter, 2024).

It has been noted in past research studies that Value alignment has become one of the major aspects of AI safety and governance, and emphasized that value alignment can be achieved only when the objectives of robotics and autonomous systems are aligned with sociocultural and ethical human values (Russell, 2019; Gabriel, 2020). This theoretical assumption is the logical antidote to the previous comparison of governance, since the diversity among regulations is a manifestation of the way society interprets the concept of “aligned values”.

Previous research studies have highlighted that “Value alignment,” as a modern concept, has emerged as one of the key components of AI safety and governance, stressing that it is possible only if the goal of Robotics is in accordance with sociocultural and ethical human values (Hadfield-Menell, 2016). This concept aligns with our research objectives to explore how ethical governance and AI literacy can enhance the shared capability to align AI technology with societal values.

Value alignment, as a modern concept, is founded on the work of Wiener, which explains how to maintain consistency in the outcomes of controlling machines, as well as on the efforts to align machines with human objectives (Muraven, 2017). Thus, value alignment works as the conceptual bridge between institutional governance models and the AI system technical design. .

The rapid progress of AI, especially generative AI, has made value alignment a core issue in governance. This is because these systems are increasingly making autonomous decisions that directly affect human welfare and social order. Frequent instances of biased, misleading, or unethical outputs from large models underscore the urgent need to ensure that AI behavior aligns with human moral and social values. The Asilomar AI Principles explicitly warn that highly autonomous systems must remain consistent with human values throughout their lifespan, reflecting a consensus among experts that "hyper-intelligent systems," if not properly managed, could deviate from ethical standards or worsen societal harms (Future of Life Institute, 2017). It is in this sense that governance structures and technical alignment mechanisms must work in a constitutive manner: institutional frameworks determine the normative boundaries, whereas technical alignment mechanisms operationalize within algorithmic systems.

The recent scholarly concept of AI value alignment is typically divided into two phases: primary alignment and secondary alignment. Primary alignment can be defined as the process of instilling human values and ethical standards directly into an AI framework during its development and training. This phase is necessary to ensure that the AI's major goals and behavioral reasoning are aligned with generally accepted human standards. Reinforcement learning based on human feedback (RLHF), incorporating ethical constraints, or introducing value-based embedding vectors are the most common methods for modelling the model’s initial decision-making structure and avoiding unacceptable results early on. Secondary alignment, in turn, is concerned with the continual adjustments to AI behavior post-implementation and its surveillance. It carries processes that allow the system to adjust its decisions and actions in response to emerging ethical issues, societal commentary, or changes in context. The focus of this stage is on constant monitoring, audit and adjustment to bring the behaviour of the AI under the changing human values and social norms as time goes by. Combined, the two phases form a dynamic framework to ensure that AI systems do not just start with values-consistent goals (primary alignment) but also maintain and strengthen that consistency throughout their operation (secondary alignment) (Huang et al., 2025).

This evolution, from static to dynamic alignment and from single authority to collaborative governance, shows that value alignment isn't a one-time coding or design project. Besides, Value alignment is constructed and co-constructed. Rather, it is an ongoing socio-technical co-construction process. Here, the previous point about multi-agent collaborative governance as an institutional coordination is furthered by stating that it is a structural requirement of enduring value congruity. At the technical level, the value alignment focuses on the systematic orientation of the ethical constraints, at the social level pertinent to negotiating the potential outcomes of agreeable AI conduct (van Wyk, 2024). Therefore, it is concluded that algorithmic optimization and inclusive governance ecosystems are equally important for effective value alignment.

Despite broad theoretical agreement on value alignment, practical implementation presents major hurdles. Examples like ChatGPT and Google Bard illustrate the challenge: it's hard to ensure that AI output consistently aligns with human ethics and cultural diversity. Since these systems can still produce biased or inappropriate content even with safety filters, it's clear that technical adjustments are insufficient (Rane

and Choudhary, 2024; Stahl and Eke, 2024; Weidinger et al., 2022). These difficulties also confirm the comparative results in Tables 1 and 2, which indicate that the regulatory design is insufficient and that adaptive monitoring and stakeholder engagement systems should be implemented.

Value pluralism and dynamic conflict: Different groups define the ethical "goodness" of AI based on their cultural values, leading to varied expectations. Western nations (Europe and the U.S.) typically view "goodness" through a human rights framework, emphasizing individual freedom, privacy, and autonomy. Meanwhile, many developing countries define it in terms of collective welfare, efficiency, and national stability, favoring AI that supports economic growth and social order (Roberts et al., 2021; NIST, 2023). This normative difference makes it difficult to establish a universally relevant standard of alignment and the argument for context-sensitive universal coordination of governance.

In this context, aligning AI with human values requires not only improvements to model design but also the establishment of comprehensive ethical evaluation systems, multi-level risk-monitoring mechanisms, and public consultation platforms. Enhancing AI literacy is a necessary foundation for developing these governance structures and provides the cognitive basis for achieving ongoing value alignment

Effective value alignment relies on the cooperation of diverse participants, including governments, technology developers, academics, and the public. Each of these groups helps translate ethical principles into practical governance through regulation, technical standards, and civic engagement. Without these collaborative mechanisms, value alignment cannot shift from an abstract idea to concrete behavioral norms this is the core meaning of multi-agent collaborative governance. The value alignment provides ethical principles aligned with societal values and sets ethical limits for AI systems. AI systems provide a platform for connecting stakeholders worldwide and serve as a space to build value consensus and ensure governance (Schwerzmann, 2025).

3. METHODOLOGY

3.1 Research Design

This paper used a Systematic Literature Review (SLR) methodology to investigate the relationship between AI literacy and ethical digital governance, and to examine how collaboration, shared values, and stakeholder involvement contribute to this relationship. The rationale for choosing the SLR approach is that it is a clear, replicable, and comprehensive system for synthesizing available evidence across disciplines (Tranfield et al., 2023).

3.2 Search Strategy

A systematic search approach was applied to large scholarly databases, including Scopus, Web of Science, IEEE Xplore, and Google Scholar, that included the literature published in 2019-2025. Search refinements were performed using Boolean operators to ensure the inclusion of relevant studies. The main keywords were used to bring together the concepts of AI literacy, ethical governance, and collaboration, e.g., (AI literacy) OR (artificial intelligence education) AND (ethical governance) OR (digital ethics) OR (responsible AI). (AI policy/AI regulation) and stakeholder participation/shared values.

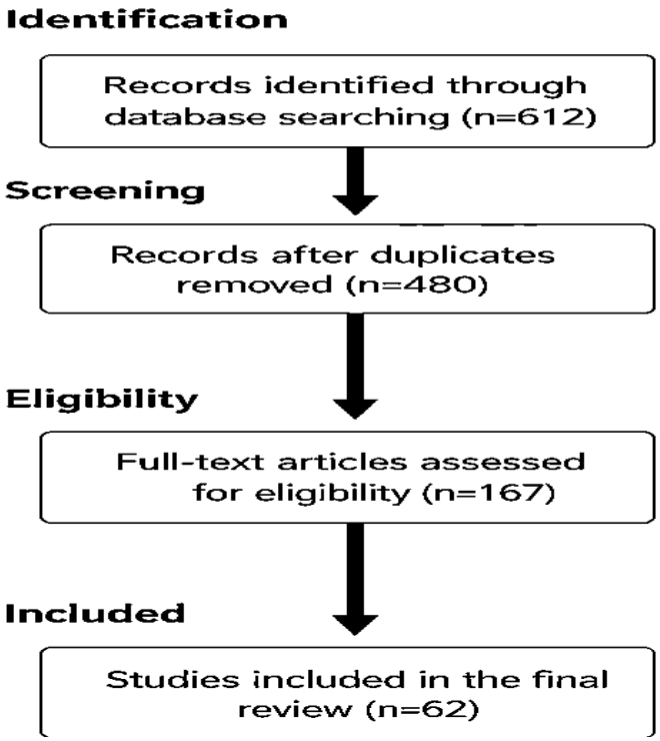
3.3 Inclusion and Exclusion Criteria

Inclusion criteria were: (1) the sources had to be peer-reviewed articles, policy reports, or institutional frameworks published in 2019-2025; (2) the source had to be in English; and (3) the source had to discuss AI governance, ethics, or literacy. Research articles that were purely technical, lacked an ethical or governance aspect, and whose complete text was not available were excluded.

3.4 Screening Process and PRISMA Flow

In accordance with the PRISMA 2020 principles (Page et al., 2021), 612 records were obtained as a result of the search process. After eliminating 132 duplicates, 480 titles and abstracts were filtered. Of the 167 full-text articles, 62 were evaluated for eligibility, and 62 studies were selected for analysis. 62 studies were included in the systematic review, while additional sources were used for conceptual and policy discussion A PRISMA flow diagram illustrated this identification, screening, and inclusion process (Figure 1).

Figure 1. PRISMA FLOW DIAGRAM



3.5 Data Extraction and Coding

Data were extracted into an ordered table that included the authors, year, geographic area of interest, main findings of each study, and the applicability of the results to AI literacy and ethical governance. The

thematic analysis was then applied to the data, following Thomas and Harden (2008). The data were manually coded to identify repetitive concepts, relationships, and policy patterns across selected studies. Thematic grouping (NVivo software) might also be applied on a large scale, but for the current study, thematic analysis was performed manually because the data were manageable and thus provided contextual sensitivity.

3.6 Thematic Analysis

The thematic analysis was performed in line with the framework of Thomas and Harden (2008), with three steps being familiarization, theme identification, and thematic synthesis. To familiarize with all 62 included studies, they were thoroughly read and summarized to identify the main trends regarding AI literacy, governance practices, and ethical implications. During the theme identification phase, common concepts emerged, including AI literacy capacity, multi-stakeholder partnerships, and value alignment across a wide variety of policy settings and scholarly discourses. Lastly, these concepts were grouped using thematic synthesis to form a coherent understanding of how AI literacy supports ethical digital governance, collaboration, and value-based decision-making (

It was discovered that three primary themes explain the process of the formation of ethical digital governance at both national and institutional levels: (1) the cognitive basis of governance is AI literacy, (2) the practical process is multi-stakeholder collaboration, and (3) the ethical result is value alignment.

3.7 Ethical Considerations

Since this was research that used secondary data (using published sources), there were no actual human subjects. Thus, the ethical permission was not needed. Nevertheless, all the examined literature and policy papers were properly referenced in order to uphold academic integrity.

4. RESULTS AND DISCUSSION

This study identified the following three main themes during the analysis, explaining how ethical digital governance develops across countries:

1. AI literacy is the cognitive foundation of governance,
2. Multi-stakeholder collaboration as the practical mechanism, and
3. Value alignment is the ethical outcome.

4.1 AI Literacy as the Foundation

The studies concluded that AI literacy goes beyond technical knowledge. It means that to understand how artificial intelligence works, we need to think critically about its social and ethical effects and be confident enough to join policy discussions or debates. Countries that invested in AI education, such as China and Singapore, through their school- and university-level curricula, reaped significant benefits. People in these countries are more informed, more engaged in ethical conversations, and less likely to be misled by misinformation about AI.

In countries that lack or have a weaker AI education policy, public understanding of AI is limited. People in these countries were less involved in shaping policies or questioning the ethical impact of emerging technologies. This shows that AI literacy builds the foundation for responsible participation in digital and good governance in the countries.

4.2 Collaboration as a Governance Mechanism

This study found that cooperation among sectors, including governments, private companies, researchers, and civil society, improves transparency and accountability in AI governance. For example, in the European Union and other OECD countries, collaborative policymaking led to stronger and clearer governance structures. The *EU AI Act* is a good example, as it brings together multiple stakeholders under a risk-based regulatory system that includes human oversight and ethical safeguards.

In a country like China, the collaboration also played an important role, primarily between the government, technology companies, and key social organizations. This created flexibility in implementation but gave the general public fewer opportunities to influence policy. The country, like the United States, encouraged open innovation and private-sector leadership; however, the lack of consistent ethical standards created inconsistencies between innovation and accountability.

4.3 Value Alignment as an Ethical Outcome

This study found that when the general public understands AI and collaborates effectively, they tend to develop shared ethical values. Systems that create dialogue among policymakers, developers, and the public are more effective in aligning AI systems with public interests. Open dialogue encourages the integration of fairness, inclusivity, and transparency into AI design and governance.

However, in culturally diverse and unequal societies, keeping value alignment is more difficult. Differences in moral perspectives and unequal access to digital tools often lead to conflicting priorities.

Overall, the study concludes that AI literacy enables individuals to participate meaningfully. At the same time, collaboration transforms ethical principles into tangible policies and institutional practices, ensuring that digital governance remains fair, inclusive, and socially responsible.

5. DISCUSSION

This work contributes to the existing knowledge of ethical digital governance by demonstrating that the concepts of AI literacy, multi-stakeholder cooperation, and value alignment can be understood as a specific ecosystem of governance rather than as separate elements. The results support and prove the theoretical views of stakeholder governance, the theory of public value, and the research on AI ethics.

First, the findings confirm the existence of a structural precondition of AI literacy to participatory governance. AI literacy is more than just coding skills; it also encompasses algorithmic awareness, critical data reasoning, and ethical reflexivity. As Ng (2021) notes, AI literacy refers to understanding, evaluating, and interacting responsibly with AI systems. Similarly, Long and Magerko (2020) define AI literacy as a skill model that enables citizens to critically evaluate the outputs of algorithms and their social consequences. These viewpoints confirm the observation that AI-conscious societies are better placed to apply democratic checks on new technologies.

Likewise, the national AI strategy of Singapore is not only technologically focused but also actively seeks to develop human capital and educational infrastructure to support the responsible Tech-Gear adoption and workforce preparation to make the country more competitive and technologically sufficient, a factor that has been classified as a priority towards organizational competitiveness (Khanal et al., 2024) China, similarly, has developed Next Generation AI Development Plan that provides a complete overhaul of the curriculum to enhance national competitiveness and technological abilities. Empirical studies have shown that these types of efforts not only make students and educators develop technical skills but also gain ethical understanding and critical thinking (Zhao et al., 2022). These reforms through education enhance the civic competence of citizens, decrease the susceptibility to misinformation, and enhance the legitimacy of AI governance by providing them with the knowledge and capacity to engage with AI technologies.

Second, the results are aligned with the Stakeholder Theory (Freeman, 1984) according to which sustainable governance is based on multi-stakeholder collaboration. The stakeholders' theory pointed out that in AI governance, stakeholders include the governments, corporations, academia, and civil society. These findings are supported by Bryson et al. (2014), who claim that the creation of public value relies on cross-sector collaboration and participatory engagement. This assertion is substantiated by the findings: jurisdictions that codify collaborative governance, such as the European Union, by the European Union Artificial Intelligence Act, have more apparent accountability frameworks and transparency processes. The EU's risk-based regulatory model is based on joint policy-making through consultation with industry experts, academic researchers, and civil society participants (European Commission, 2021). Conversely, the fragmented U.S. governance environment is more flexible regarding innovation, yet coordination challenges arise in examining federal AI regulation (Calo, 2017). These differences underscore the fact that the effectiveness of governance is not limited to the content of regulation, but also to institutionalized cooperation.

Third, the study validates that value alignment is a dynamic ethical outcome that is influenced by literacy and collaboration. According to Russell (2019), value alignment is the process of ensuring that AI systems are working according to societal norms and human preferences. Nevertheless, alignment is not fixed but it changes in response to changes in social demands and technological capabilities. According to Dignum (2019), responsible AI entails ongoing dialogue among the parties and negotiation over acceptable levels of fairness and accountability.

The argument about biased algorithmic outputs and real-world harms further supports it. Studies by Bender et al. (2021) show that large language models can replicate societal biases despite technical precautions. Such results demonstrate that technical optimization alone is not enough. Ongoing monitoring, ethical reviewing, and civil education are needed to maintain harmonization.

These findings also reflect the Public Value Theory (Moore, 1995) that holds that governance legitimacy is created by creating collective value by transparency, trust, and participation. The adoption of AI governance, integrated with literacy and collaboration, reduces the lack of trust among the population by allowing citizens to shape technological systems. The same pillars of trustworthy AI, such as inclusive growth, transparency, and accountability, are highlighted in the OECD AI policy frameworks (OECD, 2019).

However, value pluralism makes it more difficult to adhere. Most Western forms of governance tend to value human rights and personal freedoms, whereas most developing situations tend to focus on communal good and economic development. Geopolitical competition and state-centered governance logics, in particular, those focusing on digital sovereignty and national security, can reduce the room for pluralistic ethical discourse and make it more difficult to draw universal standards of AI ethics (Coeckelbergh, 2025).

The results indicate that the connection between collaboration and value alignment is moderated by AI literacy. The result highlighted that collaborative governance processes will yield legitimate and socially responsive results when citizens have more knowledge of AI. On the other hand, low literacy levels hinder meaningful interactions and undermine accountability systems. When people are well educated about AI, they are in a better position to support those policies that are developed to make the technology safer and more ethical. Also, these citizens will support efforts to point out and discuss virtual issues (Ng, 2021). This AI literacy shifts the governance from a top-down approach to a more open, participatory, and shared one.

The findings are supported by Yotawut (2018), who noted that this type of governance not only promotes efficiency but also participation, openness, and trust, which are the main themes of Public Value Theory. Coordination among institutions and the public, with proper literacy, improved the use of AI systems to enhance citizens' well-being. Such as the OECD AI Principles and the EU AI Act highlighted that collaboration among stakeholders like governments, corporates, and civil societies produced transparency, moral ethics and strengthened the governance systems (OECD, 2024). Similarly, from the comparative analysis, it was concluded that collaboration among stakeholders links AI literacy with moral outcomes. Open communication among stakeholders and public participation are the key elements of promoting shared values, building trust and accountability in the European countries. However, in developing nations, despite strong policies and less participatory strategies hindering public say in decisions and weaker ethical alignments, these findings are supported by Dignum (2019), who reported that coordination among stakeholders and public participation promote strong digital governance and ethical integrity. Furthermore, the results also highlighted that value alignment is not static but a dynamic process. The value alignment changes as people become more AI literate and technology advances. These results are supported by Biagini (2025), who pointed out that AI literacy helped promote a sense of responsibility and fairness and reduced the gap between technology and societal expectations.

.

To summarize the study results, the study contributed to the literature on AI governance by introducing a conceptualization of ethical digital governance in three layers: literacy (cognitive background), collaboration (institutional process), and value alignment (ethical product). When all three layers operate synergistically, they create a flexible and fair system of digital governance, founded on emerging technologies while keeping the public's interests at the center.

6. CONCLUSION

This paper has examined the connection between AI literacy and ethical digital governance and the role of collaboration, shared values, and stakeholder engagement in responsible technological development. Through a systematic review of academic literature, policy frameworks, and international governance models, the study established that countries vary in their approaches but share similar ethical goals. Three key themes emerged, including AI literacy as the cognitive basis of governance, multi-stakeholder collaboration as the practical process, and value alignment as the ethical impact. The results highlight the importance of AI literacy, which goes beyond technical literacy; it encompasses citizens' critical awareness of AI's social and ethical impacts. Those countries that invest in AI education, like China and Singapore, are more engaged with AI, have better policymaking, and fewer cases of misinformation. Conversely, countries with less developed AI education policies have difficulty engaging people in the system and achieving greater ethical responsibility, suggesting that literacy should be the initial step in responsible governance. The research also found that the aspect of collaborative governance enhances transparency, accountability, and trust among the people. Policies such as the EU AI Act and the OECD AI Principles can serve as examples of how governments, corporations, and civil society can together establish equitable and responsible systems of governance. Nevertheless, difficulties remain, particularly when coordination mechanisms are poor or when coordination relies on a top-down strategy. Teamwork will translate ethical issues into practical structures and is therefore an important step toward sustainable AI governance. Lastly, value alignment was seen as an active ethical process and not an objective. Stakeholder engagement in policy formulation and management strengthens the shared values and moral principles of AI systems, including inclusiveness, fairness, and human dignity. Nevertheless, balancing the culturally diverse and resource-constrained societies is not easy due to differences in priorities and access to technology. Finally, the research shows that AI literacy among citizens enhances their power, institutionalizes ethics through collaboration, and strengthens the integrity of digital governance through value alignment. These dimensions form a powerful construct for developing just, open, and humanistic AI systems capable of responding rapidly to rapid shifts in technology and society.

7. THEORETICAL IMPLICATIONS

The research presents strong theoretical contributions by synthesizing Stakeholder Theory and Public Value Theory to define ethical digital governance as a dynamic, co-constructed process. The research contributed to Stakeholder Theory further by situating AI literacy as an essential facilitator of participatory governance, focusing on informed involvement across societal and institutional levels. At the same time, it advances Public Value Theory by positing multi-stakeholder collaboration as a means of co-producing societal trust, transparency, and moral legitimacy in AI systems. The literacy collaboration value alignment model proposed here offers a coherent theoretical perspective that brings together cognitive, structural, and normative aspects of governance, thereby broadening existing explanations of how ethical, human-oriented AI governance can be transformed through participatory, literacy-based, and value-consistent frameworks.

8. PRACTICAL IMPLICATIONS

The research findings have several actionable implications. Policymakers would need to incorporate AI literacy into school, college, and university-level education curricula in every nation to foster ethical sensitivity and enable citizens to become responsible stakeholders in governance. Governments and industries would need to create multi-stakeholder forums where regulators, developers, educators, and members of the public engage in dialogue, encouraging co-regulation and accountability. Institutions would need to deploy algorithmic openness and ethical auditing mechanisms to monitor continuously and align with values. In developing countries, community-based AI education programs can bridge the digital divide and enable inclusive engagement. Lastly, ethical considerations should be integrated throughout the AI lifecycle, from design to deployment, so that technological progress supports human well-being. Together, these measures operationalize the "AI literacy collaboration value alignment" model into governance practices, reinforcing trust, transparency, and collective accountability in the AI age.

9. LIMITATIONS AND FUTURE STUDIES

Although this research will be useful in understanding the nexus between AI literacy and ethical governance, it has a number of limitations that cannot be ignored. To start with, being a systematic literature review, it used only secondary data, which can be a drawback in terms of context and accuracy of interpretation. Second, differences in the accessibility and quality of local literature, especially in the Global South, could have contributed to the comparative results. Third, the study was limited to sources published between 2019 and 2025 and may have missed earlier foundational research. Finally, the thematic synthesis has been performed manually, which, though subject to context, can be subject to subjective bias. Further studies should include empirical case studies and cross-regional surveys to substantiate and extend these conceptual insights.

Author Contributions: Xu Hao conceptualized the study, conducted the research, and drafted the manuscript. Nurul Liyana Mohd Kamil supervised the methodological design, refined the article structure, and validated the references. All authors have read and approved the final manuscript.
Ethics Statement: This study did not involve human participants or animals and therefore did not require ethical approval.
Conflict of Interest Statement: The authors declare no conflict of interest.
Data Availability Statement: The data supporting the findings of this study are available from the corresponding author upon reasonable request.
Generative Artificial Intelligence Statement: The authors declare that no generative artificial intelligence tools were used in the conceptualization, analysis, or writing of this manuscript, except for minor language editing.

REFERENCES

Alharbi, T. (2021). Assessment of cybersecurity awareness among students: Implications for protecting cyberspace and reducing cybersecurity threats. *Journal of Risk and Financial Management, 14*(2), 23. https://doi.org/10.3390/jrfm14020023

Atwood, B. (2025). Artificial intelligence in Iran: National narratives and material realities. *Iranian Studies.* Advance online publication. https://doi.org/10.1017/irn.2024.63

Batool, A., Zowghi, D., & Bano, M. (2025). AI governance: A systematic literature review. *AI and Ethics, 5*, 3265–3279. https://doi.org/10.1007/s43681-024-00653-w

Baum, S. D. (2020). Social choice ethics in artificial intelligence. *AI and Society, 35*(1), 165–176.

Beer, D. (2017). The social power of algorithms. *Information, Communication & Society, 20*(1), 1–13. https://doi.org/10.1080/1369118X.2016.1216147

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)* (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Biagini, G. (2025). Towards an AI literate future: A systematic literature review exploring education, ethics, and applications. *International Journal of Artificial Intelligence in Education, 35*, 2616–2666. https://doi.org/10.1007/s40593-025-00466-w

Birdayanthi, B., Yusriadi, Y., & Ikmal, I. (2025). Accountability and transparency in public administration for improved service delivery. *Journal Social Civilecial*. https://doi.org/10.71435/610633

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., *et al.* (2022). *On the opportunities and risks of foundation models*. arXiv. https://doi.org/10.48550/arXiv.2108.07258

Bozeman, B. (2007). *Public values and public interest: Counterbalancing economic individualism.* Georgetown University Press. https://dx.doi.org/10.1353/book13027

Bridoux, F., & Stoelhorst, J. W. (2022). Stakeholder governance: Solving the collective action problems in joint value creation. *Academy of Management Review, 47*, 214–236. https://doi.org/10.5465/amr.2019.0441

Bryson, J. M., Crosby, B. C., & Bloomberg, L. L. (2014). Public value governance: Moving beyond traditional public administration and the new public management. *Public Administration Review, 74*(4), 445–456. https://doi.org/10.1111/puar.12238

Calo, R. (2017). Artificial intelligence policy: A primer and roadmap. *UC Davis Law Review, 51*, 399–435. https://doi.org/10.2139/ssrn.3015350

Chu-Ke, C., & Dong, Y. (2024). Misinformation and literacies in the era of generative artificial intelligence: A brief overview and a call for future research. *Emerging Media, 2*(1), 70–85. https://doi.org/10.1177/27523543241240285

Coeckelbergh, M. (2025). Three challenges for a global AI ethics: Towards a more relational normative vision. *AI Ethics, 5*, 5527–5533. https://doi.org/10.1007/s43681-025-00791-9

Council of the European Union. (2024). *Artificial intelligence act*. https://www.consilium.europa.eu/en/policies/artificial-intelligence

Crawford, K., & Calo, R. (2016). There is a blind spot in AI research. *Nature, 538*(7625), 311–313. https://doi.org/10.1038/538311a

Dignum, V. (2019). *Responsible artificial intelligence: How to develop and use AI in a responsible way.* Springer. https://doi.org/10.1007/978-3-030-30371-6

Donaldson, T., & Preston, L. E. (1995). The stakeholder theory of the corporation: Concepts, evidence, and implications. *Academy of Management Review, 20*(1), 65–91.

Ebers, M. (2025). Truly risk based regulation of artificial intelligence: How to implement the EU's AI Act. *European Journal of Risk Regulation, 16*(2), 684–703.

European Commission. (2021). *Proposal for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).* Brussels.

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review, 1*(1). https://doi.org/10.1162/99608f92.8cd550d1

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Foka, A., Griffin, G., Ortiz Pablo, D., Rajkowska, P., & Badri, S. (2025). Tracing the bias loop: AI, cultural heritage and bias mitigating in practice. *AI & Society, 40*(3), 5835–5847. https://doi.org/10.1007/s00146-025-02349-z

Freeman, R. E. (1984). *Strategic management: A stakeholder approach.* Pitman.

Future of Life Institute. (2017). *Asilomar AI principles*. https://futureoflife.org/ai-principles

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines, 30*(3), 411–437. https://doi.org/10.1007/s11023-020-09539-2

Güngör, H. (2020). Creating value with artificial intelligence: A multi-stakeholder perspective. *Journal of Creating Value, 6*(1), 72–85.

Hadfield-Menell, D., Russell, S., Abbeel, P., & Dragan, A. (2016). Cooperative inverse reinforcement learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 29, pp. 3909–3917). Curran Associates, Inc

Hidayat, W., & Muis, A. (2025). Ethical and legal challenges of artificial intelligence in the judicial system: An Indonesian perspective. *Justicia Insight, 2*(1), 9–15. https://doi.org/10.70716/justin.v2i1.274

Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. *Science, 361*(6404), 751–752.

Hingle, A., & Johri, A. (2025). *Systematic review of collaborative learning activities for promoting AI literacy* (arXiv:2508.15111). arXiv. https://doi.org/10.48550/arXiv.2508.15111

Hristovska, A. (2023). Fostering media literacy in the age of AI: Examining the impact on digital citizenship and ethical decision making. *Kairos, 2*(2), 39–59. https://doaj.org/article/744cb4faf2364808b8623d9a6f750ffa

Huang, L. T. L., Papyshev, G., & Wong, J. K. (2025). Democratizing value alignment: From authoritarian to democratic AI ethics. *AI Ethics, 5*, 11–18. https://doi.org/10.1007/s43681-024-00624-1

Hussein, H., Gordon, M., Hodgkinson, C., Foreman, R., & Wagad, S. (2025). ChatGPT's impact across sectors: A systematic review of key themes and challenges. *Big Data and Cognitive Computing, 9*(3), 56. https://doi.org/10.3390/bdcc9030056

Irawati, S., Hayat, A., Juniar, A., & Handayani, S. A. (2023). Exploring accountability and transparency in government agency management: A literature review. *Ilomata International Journal of Management, 5*(3), 587–600. https://doi.org/10.61194/ijjm.v5i3.1189

Islam Tonmoy, S. M. T., Zaman, S. M. M., Jain, V., Rawte, V., Chadha, A., & Das, A. (2024). *A comprehensive survey of hallucination mitigation techniques in large language models* (arXiv:2401.01313). arXiv. https://doi.org/10.48550/arXiv.2401.01313

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., … Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), 1–38. https://doi.org/10.1145/3571730

Jiang, Z., Abedin, B., & Marjanovic, O. (2024). Understanding the components of AI literacy at the individual, group, and organisational level: An organisational learning perspective. In *Proceedings of the 35th Australasian Conference on Information Systems (ACIS 2024)*. Australasian Association for Information Systems.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399.

Karimov, A., & Saarela, M. (2025). AI literacy and governance as foundations for ethical AI: A cross-national review of government strategies. In *Proceedings of the 2025 International Conference on Innovation and Technology Research (CITREx)*. IEEE. https://doi.org/10.1109/CITREx64975.2025.10974932

Karmakar, A. (2024). AI governance in Bangladesh: A critical evaluation of the transition from strategic vision to policy framework. *Indian Journal of Legal Review, 4*(3), 668–677.

Kathala, K. C. R., & Palakurthi, S. (2025). AI literacy framework and strategies for implementation in developing nations. In *Proceedings of the 2024 16th International Conference on Education Technology and Computers (ICETC '24)* (pp. 418–422). Association for Computing Machinery. https://doi.org/10.1145/3702163.3702449

Khanal, S., Zhang, H., & Taeihagh, A. (2024). Building an AI ecosystem in a small nation: Lessons from Singapore's journey to the forefront of AI. *Humanities and Social Sciences Communications, 11*, 866. https://doi.org/10.1057/s41599-024-03289-7

Kim, T. W., Hooker, J., & Donaldson, T. (2021). Taking principles seriously: A hybrid approach to value alignment in artificial intelligence. *Journal of Artificial Intelligence Research, 70*, 871–890. https://doi.org/10.1613/jair.1.12558

Klenk, M. (2023). Algorithmic transparency and manipulation. *Philosophy & Technology, 36*(4), Article 79. https://doi.org/10.1007/s13347-023-00678-9

Kraus, S., Jones, P., Kailer, N., Weinmann, A., Chaparro-Banegas, N., & Roig-Tierno, N. (2021). Digital transformation: An overview of the current state of the art of research. *SAGE Open, 11*(3), 21582440211047576. https://doi.org/10.1177/21582440211047576

Lakshitha, P. C., Manoj, A., & Jeevanandhan, L. (2025). Artificial intelligence in Indian governance: Modernising public service delivery. *Electronic Journal of Social and Strategic Studies, 6*(Special Issue VII), 122–134. https://doi.org/10.47362/EJSSS.2025.6607

Liu, X., Zhang, L., & Wei, X. (2025). Generative artificial intelligence literacy: Scale development and its effect on job performance. *Behavioral Sciences, 15*(6), Article 811. https://doi.org/10.3390/bs15060811

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). https://doi.org/10.1145/3313831.3376727

MacDonald, A., Clarke, A., Huang, L., & Seitanidi, M. M. (2019). Partner strategic capabilities for capturing value from sustainability-focused multi-stakeholder partnerships. *Sustainability, 11*(3), 557. https://doi.org/10.3390/su11030557

Marzdar, M. H. (2025). Artificial intelligence in Iran's public administration: Opportunities, challenges, and strategic approaches for governance innovation. *International Journal of Applied Research in Management, Economics and Accounting, 2*(2), 16–35. https://doi.org/10.63053/ijmea.38

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys, 54*(6), 1–35. https://doi.org/10.1145/3457607

Miao, F., Shiohira, K., & Lao, N. (2023). *AI competency framework for students*. UNESCO. https://doi.org/10.54675/JKJB9835

Milano, S., Thiebes, S., & Avital, M. (2020). Artificial intelligence and mass personalization of communication content. *New Media & Society, 24*(5), 1258–1277. https://doi.org/10.1177/14614448221087864

Mills, K., Ruiz, P., Lee, K., Coenraad, M., Fusco, J., Roschelle, J., & Weisgrau, J. (2024, May). *AI literacy: A framework to understand, evaluate, and use emerging technology*. Digital Promise. https://doi.org/10.51388/20.500.12265/218

Moore, M. H. (1995). *Creating public value: Strategic management in government.* Harvard University Press.

Moore, M. H. (2013). *Recognizing public value.* Harvard University Press.

Muhammad, A., Siddiqui, R., & Khan, S. (2025). Ethical and governance implications of AI in Pakistan's financial sector. *Review of Management and Social Sciences, 10*(1), 45–61. https://rjmssjournal.com/index.php/7/article/view/290

Muraven, M. (2017, March 18). *Goal conflict in designing an autonomous artificial system* (arXiv:1703.06354). arXiv. https://doi.org/10.48550/arXiv.1703.06354

National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1

Ng, D. T. K. (2021). AI literacy: Definition, teaching, and learning. *Computers and Education: Artificial Intelligence, 2*, Article 100041. https://doi.org/10.1016/j.caeai.2021.100041

Ng, D. T. K., Leung, J. K. L., Chu, K. W. S., & Qiao, M. S. (2021). AI literacy: Definition, teaching, evaluation and ethical issues. *Proceedings of the Association for Information Science and Technology, 58*(1), 504–509. https://doi.org/10.1002/pra2.487

Novelli, C., Taddeo, M., & Floridi, L. (2023). Accountability in artificial intelligence: What it is and how it works. *AI & Society, 39*, 1871–1882. https://doi.org/10.1007/s00146-023-01635-y

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

Organisation for Economic Co-operation and Development. (2019). *OECD principles on artificial intelligence*. https://oecd.ai/en/ai-principles

Organisation for Economic Co-operation and Development. (2024). *OECD AI principles: OECD Council recommendation on artificial intelligence*. https://oecd.ai/ai-principles

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

Pauwels, E. (2020). *The new geopolitics of converging risks: The UN and prevention in the era of AI*. United Nations University Centre for Policy Research. https://collections.unu.edu/view/UNU:7442

Pies, I., & Valentinov, V. (2023). Trade-offs in stakeholder theory: An ordonomic perspective. *Social Responsibility Journal, 20*(5), 975–997. https://doi.org/10.1108/SRJ-06-2023-0321

Pinch, T. J., & Bijker, W. E. (1984). The social construction of facts and artifacts: Or how the sociology of science and the sociology of technology might benefit each other. *Social Studies of Science, 14*(3), 399–441. https://doi.org/10.1177/030631284014003004

Prajescu, A. I., & Confalonieri, R. (2025). *Argumentation-based explainability for legal AI: Comparative and regulatory perspectives* (arXiv:2510.11079). arXiv. https://doi.org/10.48550/arXiv.2510.11079

Prikshat, V., Patel, P., Varma, A., & Ishizaka, A. (2022). A multi-stakeholder ethical framework for AI-augmented HRM. *International Journal of Manpower, 43*(1), 226–250. https://doi.org/10.1108/IJM-03-2021-0118

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine, 380*(14), 1347–1358. https://doi.org/10.1056/NEJMra1814259

Rane, N., Shirke, S., Choudhary, S. P., & Rane, J. (2024). Education strategies for promoting academic integrity in the era of artificial intelligence and ChatGPT: Ethical considerations, challenges,

policies, and future directions. *Journal of ELT Studies, 1*(1), 36–59. https://doi.org/10.48185/jes.v1i1.1314

Rawte, V., Sheth, A., & Das, A. (2023). *A survey of hallucination in large foundation models* (arXiv:2309.05922). arXiv. https://doi.org/10.48550/arXiv.2309.05922

Roberts, H., Cowls, J., Morley, J., Taddeo, M., Wang, V., & Floridi, L. (2021). The Chinese approach to artificial intelligence: An analysis of policy, ethics, and regulation. *AI & Society, 36*, 59–77. https://doi.org/10.1007/s00146-020-00992-2

Robles, P., & Mallinson, D. J. (2023). Catching up with AI: Pushing toward a cohesive governance framework. *Politics & Policy, 51*(3), 355–372. https://doi.org/10.1111/polp.12529

Romanishyn, A. (2025). AI-driven disinformation: Policy recommendations for addressing online manipulation. *Frontiers in Artificial Intelligence, 4*, Article 1569115. https://doi.org/10.3389/frai.2025.1569115

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control.* Viking.

Russell, S., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson.

Russell, S., Dewey, D., & Tegmark, M. (2015). Research priorities for robust and beneficial artificial intelligence. *AI Magazine, 36*(4), 105–114. https://doi.org/10.1609/aimag.v36i4.2577

Sabatier, P. A., & Weible, C. M. (2014). *Theories of the policy process* (3rd ed.). Westview Press.

Sadat, A. (2025). Digital governance and civic inclusion to enhance public participation in political decision-making processes. *Frontiers in Political Science, 7*, Article 1671373. https://doi.org/10.3389/fpos.2025.1671373

Saeidnia, H. R., Hosseini, E., Lund, B. D., Alipour Tehrani, M., Zaker, S., & Molaei, S. (2025). Artificial intelligence in the battle against disinformation and misinformation: A systematic review of challenges and approaches. *Knowledge and Information Systems, 67*, 3139–3158. https://doi.org/10.1007/s10115-024-02337-7

Schwerzmann, K., & Campolo, A. (2025). "Desired behaviors": Alignment and the emergence of a machine learning ethics. *AI & Society, 40*(7), 5181–5194. https://doi.org/10.1007/s00146-025-02272-3

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (FAccT '19)* (pp. 59–68). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287598

Socol De La Osa, D. U., & Remolina Leon, N. (2024). Artificial intelligence at the bench: Legal and ethical challenges of informing—or misinforming—judicial decision-making through generative AI. *Data and Policy, 6*, 1–30. https://ink.library.smu.edu.sg/sol_research/4548

Stahl, B. C., & Eke, D. (2024). The ethics of ChatGPT: Exploring the ethical issues of an emerging technology. *International Journal of Information Management, 74*, Article 102700. https://doi.org/10.1016/j.ijinfomgt.2023.102700

State Council Information Office of the People's Republic of China. (2025). *Global AI governance action plan (full text)*. https://dusseldorf.china-consulate.gov.cn/det/zgyw/20250730_11679598.htm

Su, J., Ng, D. T. K., & Chu, S. K. W. (2023). Artificial intelligence (AI) literacy in early childhood education: The challenges and opportunities. *Computers and Education: Artificial Intelligence, 4*, Article 100124. https://doi.org/10.1016/j.caeai.2023.100124

Tahaei, M., Constantinides, M., Quercia, D., & Muller, M. (2023). *A systematic literature review of human-centered, ethical, and responsible AI* (arXiv:2302.05284). arXiv. https://doi.org/10.48550/arXiv.2302.05284

Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology, 8*(1), 45. https://doi.org/10.1186/1471-2288-8-45

Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British Journal of Management, 14*(3), 207–222. https://doi.org/10.1111/1467-8551.00375

UNESCO. (2023). *AI and education: Guidance for policy-makers in Asia-Pacific.* UNESCO Publishing.

UNESCO. (2024, August 5). *Key Nepali stakeholders provide recommendations and directions for integrating AI in education in Nepal*. https://www.unesco.org/en/articles/key-nepali-stakeholders-provide-recommendations-and-directions-integrating-ai-education-nepal

United Nations. (2022). *Global digital compact: Open consultation brief*. United Nations Office of the Secretary-General's Envoy on Technology. https://www.un.org/techenvoy/global-digital-compact

Van Wyk, B. (2024). Exploring the philosophy and practice of AI literacy in higher education in the Global South: A scoping review. *Cybrarians Journal, (73)*, 1–21. https://doi.org/10.70000/cj.2024.73.601

Veale, M., & Zuiderveen Borgesius, F. (2021). Demystifying the draft EU artificial intelligence act—Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International, 22*(4), 97–112.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Isaac, W., Haas, J., Rimell, L., & Gabriel, I. (2022). *Ethical and social risks of harm from language models* (arXiv:2112.04359). arXiv. https://doi.org/10.48550/arXiv.2112.04359

Winfield, A. F., & Jirotka, M. (2018). Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philosophical Transactions of the Royal Society A, 376*(2133), Article 20180085. https://doi.org/10.1098/rsta.2018.0085

Wright, J. (2024). The development of AI ethics in Japan: Ethics-washing Society 5.0? *East Asian Science, Technology and Society: An International Journal, 18*(2), 117–134. https://doi.org/10.1080/18752160.2023.2275987

Wu, H., Li, D., & Mo, X. (2025). Understanding GAI risk awareness among higher vocational education students: An AI literacy perspective. *Education and Information Technologies, 30*, 14273–14304. https://doi.org/10.1007/s10639-024-13312-8

Xu, J., Lee, T., & Goggin, G. (2024). AI governance in Asia: Policies, praxis and approaches. *Communication Research and Practice, 10*(3), 275–287. https://doi.org/10.1080/22041451.2024.2391204

Yotawut, M. (2018). Examining progress in research on public value. *Kasetsart Journal of Social Sciences, 39*(1), 168–173. https://doi.org/10.1016/j.kjss.2017.12.005

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education—Where are the educators? *International Journal of Educational Technology in Higher Education, 16*(1), 1–27. https://doi.org/10.1186/s41239-019-0171-0

Zhao, L., Wu, X., & Luo, H. (2022). Developing AI literacy for primary and middle school teachers in China: Based on a structural equation modeling analysis. *Sustainability, 14*(21), 14549. https://doi.org/10.3390/su142114549

Zhu, J. (2024). AI ethics with Chinese characteristics? Concerns and preferred solutions in Chinese academia. *AI & Society, 39*, 1261–1274. https://doi.org/10.1007/s00146-022-01578-w

Zou, J., & Schiebinger, L. (2018). AI can be sexist and racist—It's time to make it fair. *Nature, 559*(7714), 324–326. https://doi.org/10.1038/d41586-018-05707-8

Zwitter, A. (2024). Cybernetic governance: Implications of technology convergence on governance convergence. *Ethics and Information Technology, 26*, 24. https://doi.org/10.1007/s10676-024-09763-9