ISSE
International Journal of Service Science, Management, Engineering, and Technology
pISSN 2964-7118 eISSN 2830-604X

ALGORITHMIC ETHICS AND FAIRNESS IN BUILDING TRUST IN AI-BASED SOCIOTECHNICAL SYSTEMS

Rafadi Khan Khayru[1], Bayar Gardi[2]

[1]Universitas Sunan Giri Surabaya, Indonesia [2]Universal Technical & Vocational Institute, Iraq

correspondence: bayargardi@gmail.com

Abstract - *Artificial intelligence increasingly influences social interactions, institutional practices, and decisions affecting individual rights and opportunities. These developments raise ethical concerns regarding responsibility, algorithmic bias, fairness, transparency, and public trust. This article examines how ethical principles and algorithmic fairness shape trust in sociotechnical AI systems and how governance and regulation translate ethical commitments into institutional accountability. An integrative literature review was conducted using interdisciplinary academic publications and relevant regulatory documents. The literature was examined through thematic comparison and interpretive synthesis. The findings indicate that trustworthy AI requires more than technical accuracy or the disclosure of algorithmic information. Ethical responsibility, fairness assessments, meaningful explanations, and opportunities to contest decisions are essential components of sociotechnical trustworthiness. Governance mechanisms, including risk assessment, algorithmic auditing, human oversight, and regulatory obligations, connect normative principles with operational responsibilities. Nevertheless, transparency and formal compliance do not automatically eliminate discrimination or institutional shortcomings. The article concludes that responsible AI governance depends on continuous alignment between ethical values, technical performance, institutional accountability, and the protection of societal interests.*

Keywords: ai-ethics, sociotechnical systems, algorithmic fairness, trustworthy artificial intelligence, algorithmic transparency, ai-governance, algorithmic accountability.

INTRODUCTION

The development of artificial intelligence (AI) has changed the relationship between technology, humans, and social life. AI-based systems are increasingly used in a wide range of activities, from information management, healthcare, education, and employment to decision-making in public and private organizations. These developments show that AI holds an increasingly significant position in social processes that previously relied heavily on human judgment (Məcidova, 2026). Orlikowski (1992) explains that technology and organizations have a reciprocal relationship, because technology is shaped through human action while also influencing patterns of relationships and institutional practices. This view provides a foundation for understanding AI as part of sociotechnical systems, namely systems that bring together technological tools, humans, rules, and social structures. Through this approach, the consequences of using AI cannot be explained solely by the capabilities of algorithms, but must consider how the technology is designed, used, and overseen in the life of society.

The use of AI in decision-making raises ethical concerns because the results of algorithmic processing can affect the rights, opportunities, and well-being of individuals and groups. Decisions regarding hiring, credit eligibility, access to services, or the distribution of information can be influenced by systems that process large amounts of data. Mittelstadt et al. (2016) explain that the problems of algorithmic ethics include inconclusive evidence, the difficulty of understanding computational processes, data bias, unfair decision outcomes, changes in social relations, and the difficulty of tracing responsibility. This explanation shows that the ethical problems of AI go beyond the question of whether a system can produce accurate predictions. A decision may have high statistical accuracy and still be problematic if it results in unfair treatment or disregards the interests of certain groups. This situation places ethics and responsibility as fundamental parts of the assessment of AI-based sociotechnical systems.

Attention to AI ethics has developed through the drafting of various principles and guidelines issued by research institutions, governments, technology companies, and international organizations. Jobin et al. (2019) mapped AI ethics guidelines and found a tendency toward convergence on five principles, namely transparency, fairness, non-maleficence, responsibility, and privacy. Although the principles used are similar, the study shows differences in how they are interpreted, the reasons for applying them, and the mechanisms for implementing them. These findings show that agreement on ethical terms does not necessarily lead to uniformity in technology development practices. Organizations may declare a commitment to fairness, yet use different assessment measures in determining which outcomes are considered fair. This difference creates a need to examine how ethical principles are translated into technical decisions, organizational procedures, and accountability mechanisms that can be examined by the relevant parties.

The relationship between ethics and trust becomes a concern as society increasingly depends on AI systems to obtain information or receive decisions that have real consequences. Trust can influence individuals' willingness to use technology, accept recommendations, and hand over part of the decision-making process to automated systems. High trust does not always indicate that a system has adequate capability or quality of governance (Zanotti, 2026). Users may place excessive trust in a system because they do not understand the limitations of AI, while other users may reject systems that have in fact met safety and reliability requirements. This situation shows the need to distinguish trust as a user attitude from trustworthiness as a characteristic of the system and the institutions that manage it. Transparency, the ability to explain decisions, and clarity of responsibility are related to this assessment. In sociotechnical systems, trust develops through usage experience, the quality of decisions, and recognition of the human capacity to oversee and question technological outputs.

The issues of bias and algorithmic fairness have received attention because AI systems can learn patterns from data generated through social relations that already contain inequality. Data on education, employment, health, and digital behavior may reflect differences in access to resources between groups. Such differences do not arise from the technology itself, but from historical and structural conditions that shape who is represented in the data, how individuals are categorized, and which outcomes are recorded as success or failure. If these patterns are used to build predictions without adequate examination, automated decisions have the potential to perpetuate unequal treatment. In this way, inequality that originates in society can be carried into AI-based sociotechnical systems and presented as the result of objective computational processes. This condition makes algorithmic bias difficult to recognize, because the resulting decisions may appear consistent and statistically sound while still reproducing disadvantages experienced by particular groups. The problem is therefore not limited to individual errors in a model, but concerns how existing social relations are reinforced when AI is used in decisions that affect rights, opportunities, and well-being.

Bias can arise from data collection, category definition, variable selection, and model development through to the use of outputs in organizational decisions. At each stage, choices are made about what is measured, how groups are defined, which variables are considered relevant, and how the results are interpreted and applied. These choices involve human judgment and institutional priorities, so bias cannot be treated as a purely technical defect that appears at only one point in the process. This situation shows that bias does not always originate from programming errors or the direct use of sensitive characteristics. Variables that appear neutral may be related to particular social identities and produce different consequences. Removing sensitive characteristics from the data may therefore be insufficient, because proxy variables can still convey information related to group identity. Assessing fairness therefore requires attention to the groups that receive benefits or harms, as well as the social processes that determine how algorithmic outputs are used. Fairness cannot be judged solely from the technical properties of a model, because the same output may have different consequences depending on the institutional setting in which it is applied. A model that is used to assist human judgment may have different effects from one whose results are treated as final decisions, and the assessment of fairness needs to take these differences in use and oversight into account.

AI ethics frameworks also require a clear connection to institutional responsibility and technology governance. Floridi and Cowls (2019) developed a five-principle framework for AI that includes beneficence, non-maleficence, autonomy, justice, and explicability. This framework shows the need to consider human values in the development and use of intelligent systems. However, the existence of principles is not sufficient to explain who should conduct oversight, how unfairness is examined, and what actions are required when harm occurs (Kasaram, 2025). Ethical principles provide a normative orientation, but without institutional mechanisms they do not specify the obligations of the parties involved or the procedures through which those obligations can be examined.

This problem becomes more complex when AI is developed by one organization, provided by another company, and used by an institution that has the authority to make decisions. Each party may control only part of the process, such as model design, the provision and maintenance of the technology, or the operational use of its outputs. The division of functions can lead to unclear responsibility if it is not accompanied by oversight procedures. When harm occurs, affected parties may find it difficult to determine who should provide an explanation, who has the authority to make corrections, and who must be held accountable. The relationship between ethical values and governance therefore needs to be clarified so that normative guidelines can be translated into institutional actions that have a basis for accountability.

The understanding gap addressed by this article lies in the relationship between ethics, trust, transparency, bias, and AI governance, which are often discussed through different scholarly approaches. Ethics studies tend to emphasize normative principles, while research on algorithmic fairness has largely developed methods for measuring and reducing bias. Meanwhile, discussions of trust focus on the relationship between users and technology, whereas regulatory studies emphasize the obligations of institutions that develop and use AI. This separation can cause sociotechnical problems to be understood through a single dimension without attention to their relationship with other dimensions. It is not yet fully clear how the principles of responsibility and fairness form the basis of trust, or how transparency can be connected to oversight mechanisms and regulation. This gap calls for a synthesis that brings ethical considerations together with social experience and institutional practice so that the relationship between AI and society can be explained more fully.

This article aims to synthesize the literature on ethics and responsibility, bias and fairness, and governance and regulation as the main dimensions in the study of the relationship between AI and society. The discussion positions trust and transparency as connecting elements between system quality, user experience, and organizational accountability. Theoretically, this article is expected to clarify the relationship between ethical principles and the social mechanisms that determine the acceptance of AI technology. Academically, the synthesis is directed at bringing together studies in the philosophy of technology, sociology, information systems, and digital governance. Practically, the discussion can provide a basis for consideration for developers, institutional users, and policymakers in assessing responsibility in the use of AI. The first research question is how do ethics, responsibility, bias, and fairness shape trust and transparency in AI-based sociotechnical systems? The second research question is how can AI governance and regulation connect ethical principles with mechanisms of accountability, oversight, and the protection of societal interests? Both questions are answered through a literature synthesis that distinguishes normative, technical, and institutional issues.

RESEARCH METHODS

This study uses an integrative literature review method with a descriptive-interpretive approach to examine the dimensions of ethics, trust, transparency, fairness, and governance in AI-based sociotechnical systems. The integrative approach was chosen because the research problem brings together normative thought, technical research on algorithms, and social and institutional studies, each of which has different source characteristics. Whittemore and Knafl (2005) explain that an integrative review allows the combination of diverse forms of evidence to produce a conceptual understanding of complex problems. An integrative review can therefore produce understanding through the combination and interpretation of various scientific ideas. The descriptive-interpretive approach is used to describe how each dimension is discussed in the literature and to interpret the relationships among those dimensions in the context of trust in AI-based sociotechnical systems.

The selection of sources was carried out by considering their suitability to the research questions, the credibility of the publisher, the clarity of argumentation, and the ability of the literature to explain the relationship between ethical principles and AI management practices. The sources prioritized were academic journal articles and conceptual works that provide explanations of algorithmic fairness, trust, transparency, and institutional oversight. Regulatory documents were used to discuss the development of governance instruments, not to replace empirical research evidence. In this way, the study distinguishes between scholarly evidence, which supports the conceptual analysis, and regulatory documents, which illustrate how ethical principles are translated into legal obligations and institutional oversight.

The analysis was conducted by grouping the literature into a normative-technical dimension and an institutional dimension, then comparing the relationships among concepts and developing an argumentative synthesis. The normative-technical dimension covers ethical principles, responsibility, bias, and algorithmic fairness, while the institutional dimension covers AI governance, regulation, accountability, and oversight. Trust and transparency were treated as connecting elements between the two dimensions. This study does not use primary data, does not conduct direct audits of AI systems, and is not claimed to be a systematic review, meta-analysis, or bibliometric analysis. Its findings should therefore be read as a conceptual synthesis of the literature, not as an empirical test of the relationships discussed.

RESULTS AND DISCUSSIONS

Ethics, Algorithmic Fairness, and the Formation of Trust in Sociotechnical Systems

Ethics in AI-based sociotechnical systems relates to the principles used to assess the appropriateness of actions, decisions, and the consequences of using technology on humans. This assessment cannot be limited to the accuracy of algorithms because AI systems operate through relationships among computational tools, users, developers, and the organizations that determine the purposes of use (Paka, 2026). Selbst et al. (2019) explain that approaches to algorithmic fairness that overemphasize technical abstraction risk overlooking the social circumstances in which the system is used. This conceptual study identifies a number of problems that arise when technical modeling separates algorithmic decisions from the institutional and social relations that shape them. These findings show that a model may satisfy a particular mathematical fairness measure and still produce consequences that are socially contested. Ethical assessment requires an understanding of the purpose of use, the affected groups, and possible changes in the distribution of opportunities after the technology is implemented.

The relationship between ethics and algorithmic fairness needs to be explained through the difference between equal treatment and fairness of decision outcomes. According to Mardikaningsih and Oluwatoyin (2023), the consistent application of algorithmic rules does not necessarily produce genuine fairness. AI systems may use uniform assessment procedures for all users, but the results do not necessarily reflect equal opportunity. This difference can arise when input data depict inequality in access to education, healthcare, or economic resources. As a conceptual example, a system that assesses a person's eligibility based on digital activity history may give an advantage to groups with better access to technology. Applying the same rules does not automatically eliminate the inequality already present in the initial

information. In such circumstances, fairness requires attention to the relationship between data characteristics and the social conditions of the individuals being assessed (Eteng, 2025). Technical measures can help identify differences in outcomes between groups, but determining whether those differences are acceptable requires consideration of the purpose of the decision, the interests of the relevant parties, and the prevailing norms.

Algorithmic bias is a problem that can arise through various stages of the development and use of AI systems. Veale and Binns (2017) explain that machine learning models can produce unfair treatment through the use of historical data that contains social inequality. Variables that directly indicate sensitive identities are not the only possible source of discrimination. Information that appears neutral may have a statistical relationship with the characteristics of particular groups and produce inequality through a proxy variable mechanism. Removing sensitive information from the data is not necessarily sufficient to address this problem. In practice, developers need to pay attention to data quality, relationships among variables, and possible differences in outcomes across user groups. Bias examination also faces difficulties when information on sensitive characteristics is unavailable or is not appropriate to collect (Swami, 2026). This situation shows a complex relationship between privacy protection and the need to identify algorithmic discrimination.

Privacy issues are part of AI ethics because system development often requires the collection and processing of information about individuals' behavior and characteristics. Nissenbaum (2004) developed the concept of contextual integrity, which explains that privacy protection relates to the appropriateness of information flows with respect to the norms prevailing in a social relationship. This perspective shows that data use cannot be assessed solely on whether the information was given voluntarily or is technically available. The use of information needs to consider the purpose of collection, the recipient, and the expectations attached to the relationship in which the data was obtained. In AI systems, information that was originally collected to provide a service can be reused to build predictions about individual behavior (Prajapati, 2025). This change of purpose can raise ethical concerns even if the computational process is accurate. Privacy protection becomes part of maintaining a relationship that respects users' rights and expectations regarding the management of their personal information.

The relationship between privacy and fairness shows that applying one ethical principle can generate requirements different from those of another principle. Examining discrimination may require information on group characteristics, whereas privacy protection encourages limits on the collection of sensitive data. This problem cannot be resolved merely by choosing to collect all information or to delete all attributes related to identity. Organizations need to consider the purpose of the examination, the possible harms, and appropriate safeguard mechanisms (Carvalho et al., 2023). For example, an evaluation of differences in system outcomes can use group data that is managed in a restricted manner with access control procedures. This option still requires an assessment of the appropriateness of the processing and the relevant consent (Darmawan, 2024). This situation shows that AI ethics involves weighing several values that may have different implementation requirements. The management of value conflicts must be explainable so that technical decisions do not conceal normative considerations that affect users' rights.

Trust in AI relates to an individual's assessment of the capability, reliability, and appropriateness of using a system to assist with or carry out a task. Lee and See (2004) explain that trust is related to a person's decision to rely on automation, especially when the system has characteristics that are difficult to understand fully. However, automated decision-making should not eliminate human involvement in overseeing and evaluating the results produced by AI (Gani & Darmawan, 2022). They emphasize the need to build a level of trust that corresponds to the actual capabilities of the technology so that users neither over-rely on nor excessively disregard the system. According to Darmawan (2025b), decisions about which types of tasks are delegated to AI systems cannot be determined solely from a technological perspective. This view provides a basis for understanding that increasing trust is not always an appropriate goal if the capabilities of AI still have limitations. In services that use automated recommendations, users need to know when the system can be relied upon and when a decision requires additional examination. Calibrated trust helps maintain the human role in using technology responsibly (Coleman, 2026). This relationship shows that clarity about capabilities and limitations is part of the quality of a trustworthy system.

The formation of trust in AI is also influenced by the character of the interaction and the representation of the technology that users encounter. Glikson and Woolley (2020) reviewed empirical research on human trust in AI and developed a framework that considers the form of AI representation and the capabilities of the system as elements that influence cognitive and emotional trust. Human responses to AI cannot always be explained using technical measures alone. Users may give different assessments to a system that produces recommendations through a conversational interface compared with a system that operates without direct interaction. Experiences of success and failure can also shape assessments of the reliability of the technology. In sociotechnical systems, this relationship is influenced by the way organizations introduce AI and explain the responsibility for its use. Trust that develops through appearance or communication needs to be distinguished from the trustworthiness of a system based on data quality, decision accuracy, and oversight mechanisms.

Transparency is an element related to trust because users and affected parties need information to understand the use of AI in decisions that affect them. Transparency is not merely an abstract moral demand, but a foundation of organizational trust (Darmawan, 2022). Transparency has different forms and limitations. Burrell (2016) distinguishes

three sources of algorithmic opacity, namely intentional secrecy by organizations, limitations of technical understanding, and complexity arising from the character of machine learning. This distinction shows that providing access to program code is not always sufficient to explain how a decision is produced. Users who lack technical knowledge may still experience difficulty even when programming information is available. Meanwhile, systems with complex mathematical relationships can produce decisions that are difficult to translate into simple explanations. Transparency needs to be designed according to the needs of the party receiving the information. The aim is to help understanding and examination of decisions, not merely to increase technical information that is difficult to use.

Algorithmic explanation is related to an individual's ability to assess a decision and take action when experiencing an adverse outcome. In public services, for example, a person who receives a rejection based on an AI recommendation needs an explanation of the factors related to that decision. An explanation of the reasons for a decision, accompanied by a mechanism for human review, can demonstrate respect for users (Darmawan, 2025a). The information provided needs to distinguish the reasons for an assessment from a mere general description of how the technology works. The explanation also needs to consider whether the recipient has the opportunity to correct erroneous information or request a re-examination (Wahab & Suhardi, 2026). Without a clear understanding of how algorithms produce decisions, any bias that may arise will be difficult to identify and correct (Radjawane & Mardikaningsih, 2022). This situation shows that transparency has informational and procedural functions. Clear information can help users understand decisions, while an objection mechanism provides an opportunity to question whether a decision conforms to the applicable rules. Without such channels, transparency may stop at the delivery of information without improving the situation of the individual who suffers harm. The quality of an explanation needs to be assessed through its benefit to the affected parties and their ability to use the information provided (Darmawan, 2024).

Buijsman (2024) developed a value-based approach to transparency that focuses on the principles chosen during system design and the success of their implementation. This approach does not only demand explanations of development procedures or test results, but links them to the values that the technology is intended to realize. Within this framework, organizations need to explain the values considered relevant, how those values are translated into requirements, the results of assessments of their fulfillment, and the conflicts that arise in decision-making. This view provides a basis for understanding transparency as a means of normative accountability. Transparency is then no longer limited to describing how a system works, but also concerns whether the system can be examined against the ethical commitments that an organization has declared.

A system that declares a commitment to fairness, for example, needs to explain the fairness measure used and the reasons for choosing it. Without such an explanation, the relevant parties cannot know what kind of fairness is being pursued, which groups are taken into account, or how conflicts between values are resolved. Transparency becomes more meaningful when information enables the relevant parties to know the relationship between ethical promises and system performance. The delivery of information still requires independent examination so that it does not depend entirely on an organization's assessment of itself. Independent examination makes it possible to check whether the information provided is accurate, complete, and consistent with the actual behavior of the system.

The relationship between ethics, trust, and transparency shows that each dimension has interrelated functions but cannot replace the others. Ethics provides the basis for assessing appropriate action, fairness relates to the distribution of benefits and harms, while transparency helps explain the process and reasons for decisions (Eddine et al., 2023). Trust develops through users' assessment of the capabilities of the technology and of the institutions that manage it. Because each dimension addresses a different aspect of AI-based sociotechnical systems, strengthening one of them does not automatically resolve the weaknesses of the others. A discussion that treats trust as the outcome of transparency alone, or of ethical statements alone, therefore overlooks the different roles that these dimensions perform.

This limitation can be seen in two contrasting situations. A system may provide explanations that are easy to understand and still produce inequality if its data or decision criteria are problematic. Conversely, a system with relatively fair assessment outcomes may have difficulty gaining acceptance if users do not understand how decisions are made. This situation shows the need for alignment between the quality of outcomes, the openness of procedures, and the experience of the affected parties. In this article's interpretation, trustworthiness should be positioned as a characteristic built through the fulfillment of obligations, not merely through the formation of positive user perceptions. Trust that is formed only through appearance or communication does not guarantee that the system and the institutions managing it deserve that trust.

Ethics, responsibility, bias, and fairness form the basis for assessing the trustworthiness of AI systems in social life. Algorithmic bias can arise through data, modeling, and institutional decisions, while fairness requires attention to the experiences and distribution of consequences across groups. Transparency helps users understand the basis of decisions and the possible limitations of the system, but its benefit depends on whether the information can be used to conduct examination. Appropriate trust is built through reliability, openness, and the opportunity to question the results of technology. Trust is the result of the relationship between technical quality and social responsibility. These principles are not sufficient to guarantee organizational behavior if they are not followed by clear oversight and accountability

mechanisms. The relationship between ethical values and institutional implementation becomes the basis for the discussion of AI governance and regulation.

AI Governance and Regulation as Mechanisms of Accountability and Social Protection

AI governance is the arrangement that determines how systems are designed, developed, used, and overseen so that they conform to organizational objectives and obligations to society. This arrangement includes the assignment of responsibility, risk assessment, information management, and mechanisms for handling decisions that cause harm (Johannes et al., 2022). Kroll et al. (2017) explain that the development of automated decision-making creates a need for accountability tools capable of examining whether algorithmic decisions conform to legal and institutional standards. They show that oversight mechanisms previously developed for human decisions are not always adequate when applied to computational systems. This problem arises because AI decisions can involve processes that are difficult to explain and a division of responsibility among different parties. Governance is needed to ensure that technological decisions can still be examined on the basis of organizational obligations. Oversight needs to be directed at development procedures and at the social consequences when the system is used.

The division of responsibility in AI systems becomes a problem because the development and use of technology often involve several parties with different authority. Developers determine the model design and data processing procedures, while technology providers may manage updates and maintenance of the tools. The user organization then sets operational objectives and uses the system's outputs to make particular decisions. The process of decision-making in automated technological systems does not eliminate the responsibility of the parties who decide to use and control the technology (Azizah et al., 2024). When a decision causes harm, this division of functions can make it difficult to determine which party must provide an explanation or make corrections (Chen, 2026). This situation shows the need to distinguish technical responsibility, management responsibility, and responsibility toward affected parties. This division does not mean that every harm must automatically be assigned equally to all parties. The determination of obligations needs to consider the level of control, the decisions made, and the standards applicable to each actor. This difference cannot be resolved through moral intuition alone, but through an assessment of the level of control, knowledge, and contribution (Sutanto et al., 2023). Governance can help clarify the relationships among parties before the system is used in activities that have social consequences (Darmawan et al., 2020).

Algorithmic accountability relates to the obligation of parties who develop or use AI to provide explanations and to accept examination of the actions and outcomes of the system. Wieringa (2020) conducted a systematic review of algorithmic accountability and showed that the concept involves various actors, standards, processes, and accountability relationships. This study shows that accountability is not sufficiently understood as the ability to explain algorithms. Algorithms become governance mechanisms that perform managerial functions, ranging from task allocation, oversight, and performance evaluation to control over individual activities (Hambodo & Darmawan, 2024). Parties with authority also need to be answerable through clear procedures. In organizations, accountability can be realized through the designation of those responsible for the system, documentation of decisions, and compliance examination against predetermined standards (Nguyen et al., 2024). The existence of documentation does not necessarily guarantee corrective action when problems occur. Accountability requires a connection between information, examination authority, and the possibility of consequences for violations. This connection helps distinguish genuine accountability from administrative activity that merely demonstrates the fulfillment of formal procedures.

Algorithmic auditing is one mechanism that can be used to examine whether AI systems conform to ethical and governance requirements. Raji et al. (2020) developed an internal auditing framework that covers the AI development process comprehensively. The framework emphasizes the need for documentation and examination at various stages so that organizations can identify possible problems before the system is deployed as well as after it is used. This approach shows that auditing should not be limited to testing the final outcomes of a model. Decisions regarding the purpose of development, data sources, the selection of success measures, and the implementation of the system also require attention. In practice, auditing can help organizations identify gaps between ethical commitments and the actions that take place during technology development. Nevertheless, the success of an audit depends on the clarity of standards, the quality of evidence, and the authority of the examining party. An audit that only produces a report without follow-up is not sufficient to ensure that social risks have been adequately addressed.

Risk assessment needs to consider all stages of the life of an AI system because problems can arise before a model produces decisions as well as after the technology is used. Suresh and Guttag (2021) developed a framework that identifies various sources of harm throughout the machine learning lifecycle, from data collection to the development and deployment of the system. This study shows that sources of harm can originate from choices regarding data, measurement, modeling, evaluation, and the use of technology in real-world environments. This understanding reveals the limitations of oversight that uses only a single measure of model accuracy. A system may show good test results and still face difficulties when used on groups or in circumstances that differ from its development data. In sociotechnical systems, changes in user behavior and institutional practices can also affect the consequences of using AI. Oversight needs to include post-deployment monitoring so that new problems can be identified and addressed through appropriate action.

The relationship between ethical principles and auditing is explained through the research of Laine et al. (2024), which found that principles such as fairness, transparency, responsibility, privacy, trust, and non-maleficence are interpreted through diverse approaches. The study also explains that auditing can contribute guidelines, examination methods, and improved understanding for stakeholders. This finding shows that ethical auditing requires clarity about the values that serve as the basis of the examination. If an organization uses the term fairness without establishing the measures and testing procedures for it, the auditing process risks producing conclusions that are difficult to compare. The examination also needs to consider the interests of users, developers, regulators, and affected society. This diversity of interests makes auditing an activity that requires normative and technical judgment at the same time.

Transparency in governance has limitations if it is understood only as the disclosure of information about design or computational processes. Ananny and Crawford (2018) criticize the assumption that the openness of a system automatically enables society to understand and oversee algorithmic decisions. They show the limitations of the transparency approach, including differences in the ability to access information and power relations that affect who can use that information. This conceptual finding shows that transparency needs to be linked to the ability to conduct examination and to influence organizational action. In AI systems used for public services, providing information about the use of algorithms may help society become aware of the existence of the technology (Lünich & Starke, 2026). That information does not necessarily enable a person to assess whether the decision they received is appropriate. Oversight requires procedures that provide opportunities to ask questions, obtain explanations, and request re-examination. Transparency becomes part of accountability when information can be used to question and correct decisions.

The development of regulation shows an effort to translate AI ethical principles into legal obligations and institutional oversight. The European Union established Regulation (EU) 2024/1689, or the Artificial Intelligence Act, in 2024 as a regulatory framework for AI that uses a risk-based approach. The regulation differentiates obligations according to the character of use and the level of risk the system poses to safety, health, and fundamental rights. AI systems categorized as high-risk are subject to requirements regarding risk management, data quality, documentation, traceability, transparency, and human oversight. This arrangement shows a shift from statements of principle toward the establishment of obligations that can be examined. The application of its provisions has different stages and effective dates according to the arrangements in the regulation. In addition, European Union regulation cannot be assumed to apply automatically in all countries. Its value for this study lies in the picture it provides of how AI risks are translated into legal responsibility for those who develop and use the technology.

Clarity about legal responsibility needs to be part of the standards of digital technology governance so that every party involved understands the obligations and consequences of using the systems they develop or manage (Rofiqi et al., 2025). The risk-based regulatory approach provides a basis for differentiating oversight needs according to the possible consequences of using AI. Systems used in activities with limited consequences may require different arrangements from technologies that affect access to employment, education, healthcare, or basic services. This distinction helps direct oversight resources to uses with a greater possibility of harm. Determining the level of risk requires careful assessment because social consequences can change when a system is applied in a different environment (Lanzolla et al., 2026). A model that was initially used to assist administrative decisions may have greater consequences if its results are treated as final decisions without human review. This situation shows that risk classification needs to consider the purpose of use and institutional practices. Oversight also requires continuous evaluation when a system is updated, uses new data, or is applied to different groups of users. Regulation becomes more meaningful when its provisions are linked to real implementation and evaluation.

The relationship between governance and trust requires an understanding of the interconnection among technology, user institutions, and the arrangements that give authority to both. Lahusen et al. (2024) developed an interdisciplinary discussion of trust, trustworthiness, and AI governance, especially in the use of algorithmic systems by public authorities. The study emphasizes that the trustworthiness of AI needs to be understood through the relationship among technological applications, administrative practices, and regulatory systems. This explanation shows that the quality of algorithms is not necessarily sufficient to build trust if the user organization does not have adequate accountability procedures. Conversely, the existence of formal rules does not necessarily guarantee trustworthiness if the system used produces inaccurate or discriminatory decisions. Trust in AI requires alignment among technical capability, institutional behavior, and oversight that can be held accountable. This relationship shows the need for cooperation among developers, institutional users, regulators, and the societal groups that experience the consequences of using the technology.

Responsible AI governance requires institutional mechanisms that can connect risk assessment with corrective action. Organizations can designate those responsible for the technology, develop documentation procedures, and examine the possibility of discrimination against particular user groups. This examination needs to be supported by a clear division of authority so that the problems found do not stop at an evaluation report. When users experience harm as a result of automated decision-making, channels must be available that allow information to be examined and corrective action to be considered. Human oversight also needs to have real authority, not merely formal involvement once the system has already determined all outcomes (Evans & Singh, 2026). In certain circumstances, a decision may

require the temporary suspension of the use of a model until the problems found are addressed. Involving affected groups can help organizations understand consequences that are not visible through technical measures. These mechanisms constitute a conceptual synthesis directed at bringing ethical principles together with operational responsibility in the management of AI.

AI governance and regulation can connect ethical principles with accountability through the assignment of responsibility, risk assessment, auditing, transparency, human oversight, and mechanisms for correcting decisions. Regulation provides the basis for obligations, while organizational arrangements determine how those obligations are applied throughout the process of developing and using the technology. Auditing helps identify discrepancies between declared values and system behavior, while public oversight provides an opportunity to question the consequences that arise. The protection of societal interests requires a connection among technical quality, institutional responsibility, and regulatory standards. This interconnection becomes the basis for forming trustworthy AI systems. The success of governance still depends on real implementation, oversight capability, and the willingness of the responsible parties to make corrections when risks or harms are found.

CONCLUSIONS

Ethics, responsibility, bias, and fairness are interrelated dimensions in the formation of trust and transparency in AI-based sociotechnical systems. Trust in technology cannot be built solely through predictive capability or computational accuracy, but requires clarity of purpose, protection of rights, examination of unfairness, and explanations of decisions that can be understood. Algorithmic bias can develop through data, modeling, and institutional practices, so addressing it requires both technical and social consideration. Meanwhile, AI governance and regulation connect ethical principles with accountability mechanisms through the establishment of obligations, auditing, human oversight, and risk management. The synthesis of the two research questions shows that trustworthy AI systems require alignment among normative values, technological quality, and organizational responsibility. Transparency and fairness acquire practical meaning when both are supported by procedures that enable society to understand, examine, and question decisions.

The results of the discussion provide theoretical implications for understanding AI as a sociotechnical system shaped through the relationship among technology, humans, and institutions. Ethics, trust, transparency, fairness, and governance need to be understood as interrelated dimensions in determining the appropriateness of using AI. Academically, this synthesis brings normative thought together with technical and social studies of algorithmic decisions. The discussion shows that assessing the success of technology requires attention to the distribution of benefits, the possibility of harm, and the ability of affected parties to obtain accountability. Practically, the results of this study provide a basis for organizations to develop oversight mechanisms appropriate to the character of AI use. Developers and institutional users need to clarify responsibility, document decisions, and provide channels for examination. These implications position public trust as the result of system quality and the fulfillment of obligations, not as a standalone communication goal.

Organizations that develop and use AI are advised to translate ethical principles into requirements that can be examined throughout the process of developing and deploying the technology. Fairness assessment needs to be carried out with attention to data quality, differences in outcomes across groups, and the possibility of harm arising from the use of the system. Transparency should be directed at providing information that can be understood by users and affected parties, accompanied by the opportunity to file objections and obtain re-examination. The division of responsibility among developers, providers, and user organizations needs to be clearly established. Future research is advised to test the relationships among transparency, experiences of fairness, quality of oversight, and trust in AI through an empirical approach. Comparisons of AI use in the education, healthcare, employment, and public service sectors are also needed to explain differences in governance needs. Longitudinal research can help in understanding how trust and accountability change when a system is updated or encounters problems during its use.

REFERENCES

Ananny, M., & K. Crawford. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media & Society, 20(3), 973–989. https://doi.org/10.1177/1461444816676645

Azizah, F., R. Mardikaningsih, E. C. Gautama, S. Sulani, & S. N. Halizah. (2024). Antitrust on dynamic pricing algorithms in e-commerce: a normative legal analysis. International Journal of Service Science, Management, Engineering, and Technology, 5(2), 42–52.

Buijsman, S. (2024). Transparency for AI systems: A value-based approach. Ethics and Information Technology, 26(34), 1–11. https://doi.org/10.1007/s10676-024-09770-w

Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. Big Data & Society, 3(1), 1–12. https://doi.org/10.1177/2053951715622512

Carvalho, T. S., N. Moniz, & L. Antunes. (2023). A three-way knot: privacy, fairness, and predictive performance dynamics. Progress in Artificial Intelligence, 55–66. https://doi.org/10.48550/arXiv.2306.15567

Chen, Z. (2026). Operational responsibility in AI governance: a user-centric liability framework. AI and Ethics, 6(3), 1–16. https://doi.org/10.1007/s43681-026-01163-7

Coleman, B. (2026). Trust in human–artificial intelligence interactions. Social Science Research Network, 1–62. https://doi.org/10.2139/ssrn.6758420

Darmawan, D. (2022). Posthuman human resource management in organizations using generative artificial intelligence. Studi Ilmu Sosial Indonesia, 2(2), 97–124.

Darmawan, D. (2024). Distribution of six major factors enhancing organizational effectiveness. Journal of Distribution Science, 22(4), 47–58.

Darmawan, D. (2025a). Customer confidence as a business asset: The role of transparency and data ethics in loyalty and reputational resilience. International Journal of Service Science, Management, Engineering, and Technology, 7(3), 35–44.

Darmawan, D. (2025b). Job redesign, automation, and digital workforce competencies in the twenty-first century. Bulletin of Science, Technology and Society, 4(3), 103–114.

Darmawan, D., R. Mardikaningsih, E. A. Sinambela, S. Arifin, A. R. Putra, M. Hariani, M. Irfan, Y. R. Al Hakim, & F. Issalillah. (2020). The quality of human resources, job performance and employee loyalty. International Journal of Psychosocial Rehabilitation, 24(3), 2580–2592.

Eddine, B. A. S., D. Darmawan, R. Mardikaningsih, & E. A. Sinambela. (2023). The effect of knowledge management and quality of work life on employee commitment. Journal of Human Sciences, 10(1), 87–100.

Eteng, S. (2025). Bias, fairness, and ethical AI. Journal of Cyber-Physical Security and Robotics, 1(1), 26–38. https://doi.org/10.64235/jocpsr.01.1.08

European Parliament & Council of the European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, 1–81. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

Evans, A., & A. Singh. (2026). Key aspects of AI governance. In Book AI Governance, 93–109. https://doi.org/10.4324/9781003682530-11

Floridi, L., & J. Cowls. (2019). A unified framework of five principles for AI in society. Harvard Data Science Review, 1(1), 535–545. https://doi.org/10.1162/99608f92.8cd550d1

Gani, A., & D. Darmawan. (2022). Ethics and accountability in artificial intelligence-based managerial decision making. Journal of Social Science Studies, 2(1), 147–152.

Glikson, E., & A. W. Woolley. (2020). Human trust in artificial intelligence: Review of empirical research. Academy of Management Annals, 14(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Hambodo, T., & D. Darmawan. (2024). Regulatory gaps and legal protection for flexible workers in the digital platform economy. Journal of Social Science Studies, 4(2), 277–290.

Jobin, A., M. Ienca, & E. Vayena. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1, 389–399. https://doi.org/10.1038/s42256-019-0088-2

Johannes, S., M. Christian, & B. J. vom. (2022). Artificial intelligence governance for businesses. Information Systems Management, 1–21. https://doi.org/10.48550/arxiv.2011.10672

Kasaram, D. (2025). Reimagining ethical foundations for responsible artificial intelligence: a comprehensive framework for future ai development, governance, and sustainable human-centered implementation. International Journal of Computer Engineering and Technology, 16(1), 4263–4289. https://doi.org/10.34218/ijcet_16_01_289

Kroll, J. A., J. Huey, S. Barocas, E. W. Felten, J. R. Reidenberg, D. G. Robinson, & H. Yu. (2017). Accountable algorithms. University of Pennsylvania Law Review, 165(3), 633–705. https://pennlawreview.com/2017/02/23/accountable-algorithms/

Lahusen, C., M. Maggetti, & M. Slavkovik. (2024). Trust, trustworthiness and AI governance. Scientific Reports, 14(1), 20752. https://doi.org/10.1038/s41598-024-71761-0

Laine, J., M. Minkkinen, & M. Mäntymäki. (2024). Ethics-based AI auditing: A systematic literature review on conceptualizations of ethical principles and knowledge contributions to stakeholders. Information & Management, 61(5), 103969. https://doi.org/10.1016/j.im.2024.103969

Lanzolla, G., M. Pagani, & C. L. Tucci, C. L. (2026). Scaling AI with adaptive governance. MIT Sloan Management Review, 1–10. https://doi.org/10.63383/lvmh6418

Lee, J. D., & K. A. See. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lünich, M., & C. Starke. (2026). The public legitimacy of artificial intelligence governance. Handbook on the Global Governance of Artificial Intelligence, 283–298. https://doi.org/10.4337/9781035338580.00025

Məcidova, L. (2026). The role of artificial intelligence in modern society and ethical problems. Elmi Əsərlər., 93(5), 47–51. https://doi.org/10.69682/arti.2026.93(5).47–51

Mardikaningsih, R., & F. Oluwatoyin. (2023). Analyzing algorithmic bias, automated justice, and social transformation in artificial intelligence implementation. Studi Ilmu Sosial Indonesia, 3(1), 107–128.

Mittelstadt, B. D., P. Allo, M. Taddeo, S. Wachter, & L. Floridi. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2), 1–21. https://doi.org/10.1177/2053951716679679

Nguyen, L. H., S. Lins, M. Renner, & A. Sunyaev. (2024). Unraveling the nuances of AI accountability: A Synthesis of dimensions across disciplines. Thirty-Second European Conference on Information Systems (ECIS 2024), 1–16. https://doi.org/10.5445/ir/1000170105

Nissenbaum, H. (2004). Privacy as contextual integrity. Washington Law Review, 79(1), 119–157. https://digitalcommons.law.uw.edu/wlr/vol79/iss1/10/

Orlikowski, W. J. (1992). The duality of technology: Rethinking the concept of technology in organizations. Organization Science, 3(3), 398–427. https://doi.org/10.1287/orsc.3.3.398

Paka, C. (2026). Operationalizing ethical AI. OSF. https://doi.org/10.17605/osf.io/yg9n4

Prajapati, S. B. (2025). Existing challenges in ethical AI. World Journal of Advanced Research and Reviews, 25(1), 2130–2137. https://doi.org/10.30574/wjarr.2025.25.1.0267

Radjawane, L. E., & R. Mardikaningsih. (2022). Building ethical and fair technology: approaches to responsible technology development and application. Journal of Social Science Studies, 2(1), 189–194.

Raji, I. D., A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, & P. Barnes. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 33–44. https://doi.org/10.1145/3351095.3372873

Rofiqi, M. A., R. K. Khayru, & D. Darmawan. (2025). Digital product liability: A legal case study of consumer losses arising from mobile applications. International Journal of Service Science, Management, Engineering, and Technology, 7(1), 17–26.

Selbst, A. D., D. boyd, S. A. Friedler, S. Venkatasubramanian, & J. Vertesi. (2019). Fairness and abstraction in sociotechnical systems. In Proceedings of the Conference on Fairness, Accountability, and Transparency, 59–68. https://doi.org/10.1145/3287560.3287598

Suresh, H., & J. Guttag. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In Proceedings of the 2021 ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization. Association for Computing Machinery, 1–9. https://doi.org/10.1145/3465416.3483305

Sutanto, H., R. Saputra, & D. Darmawan. (2023). Legal violation patterns in digital technology on liability and proof. Studi Ilmu Sosial Indonesia, 3(2), 277–308.

Swami, P. P. (2026). Bias detection and fairness improvement in machine learning models. International Journal of Artificial Intelligence and Machine Learning, 6(10), 1923–1929. https://doi.org/10.51483/ijaiml.6.10s.2026.1923-1929

Veale, M., & R. Binns. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. Big Data & Society, 4(2), 1–17. https://doi.org/10.1177/2053951717743530

Wahab, A., & M. Suhardi. (2026). Algorithmic transparency in digital public services: an administrative law perspective. Jurnal Kecerdasan Buatan dan Teknologi Informasi, 5(1), 93–100. https://doi.org/10.69916/jkbti.v5i1.461

Whittemore, R., & K. Knafl. (2005). The integrative review: Updated methodology. Journal of Advanced Nursing, 52(5), 546–553. https://doi.org/10.1111/j.1365-2648.2005.03621.x

Wieringa, M. (2020). What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 1–18. https://doi.org/10.1145/3351095.3372833

Zanotti, G. (2025). AI systems should be trustworthy, not trusted. Ai & Society, 41(4), 3401–3412. https://doi.org/10.1007/s00146-025-02728-6