

Ethical Analysis in the Development of AI Systems that Automate Decision Making

Eric Krisna Sandi

Universitas Nurdin Hamzah

erickrisna1@gmail.com

Abstract

The development of artificial intelligence (AI) systems capable of making automatic decisions raises complex ethical challenges that require thorough examination. This study aims to explore how ethical principles such as fairness, transparency, responsibility, and accountability are operationalized in the AI development process. Using a qualitative approach with a multiple case study design, data were collected through literature review, semi-structured interviews with developers, ethicists, and policymakers, as well as analysis of internal policy documents and algorithm audits. The findings indicate that although there is awareness of the need to apply ethical principles, the integration of ethics into the technology development cycle remains incomplete. Key obstacles include the lack of harmonized audit standards, limited organizational capacity to foster an ethical culture, and insufficient technical tools to translate abstract principles into concrete practice. Technical approaches, such as the use of synthetic data and the “Ethics as a Service” governance model, have proven effective in enhancing fairness and accountability, yet their success depends heavily on institutional support and ongoing evaluation. This study recommends the development of a contextual and adaptive ethical framework for AI advancement in Indonesia.

Keywords: Ethical Analysis; AI Systems; Automate Decision Making

Introduction

The development of artificial intelligence technology has progressed at a rapid pace in recent years across the globe. Today, it serves as more than just a tool; it is also used to make automated decisions across a wide range of sectors, including healthcare, driverless transportation, financial systems, and legal and security services. This shows that artificial intelligence is not just a tool, but also plays an important role in decision-making that can affect people's lives.

In many countries, the issue of injustice is increasingly being raised due to biases in data, closed decision-making processes, and the reduced role of humans in dealing with the impact of automated decisions determined by AI systems. Bias in data can lead to discrimination, for example when algorithms favor one particular group, while non-transparent processes make it impossible for users to know the reasoning behind each decision, making it difficult to determine who should be held accountable. Without clear transparency and accountability, artificial intelligence can operate without effective oversight, which can threaten social justice and undermine public trust. Likewise, facial recognition technology has been shown to produce inaccurate results when recognizing individuals with dark skin, which can lead to misidentification and unfounded arrests.

Transparency in decision-making processes faces a major challenge as many AI systems function as black boxes. As a result, users or affected parties are unaware of how and why a decision was made, making it difficult for them to ask for explanations or clarify the results set by the system. Moreover, transparency is essential for building trust; without clear explanations, people are less likely to trust AI, and this can lead to resistance to the use of algorithms, especially in high-risk situations.

Research has also shown that transparency, both in the form of access to the model logic and the system's ability to explain its results, can increase perceptions of trust and allow users to check for possible biases or errors. However, achieving transparency is not easy as model complexity, intellectual property protection, and the need to keep data secure are key challenges. Therefore, solutions such as Explainable AI, external audits, and user education should be implemented to make AI operate in a more open, understandable, and accountable manner.

Floridi and his colleagues emphasize three main ethical principles that form the basis for responsible AI development:

1. Fairness - to avoid algorithmic discrimination and ensure equal access and treatment for all groups.
2. Transparency - so that decision-making processes can be audited, understood and improved as needed.
3. Accountability - to ensure that humans retain ultimate control over the decisions generated by the system. Without these three principles, artificial intelligence can exacerbate existing injustices, undermine public trust, and turn into a tool of control instead of empowerment of human rights.

The government, through the Agency for the Assessment and Application of Technology, has designed the National Strategy for Artificial Intelligence 2020-2045 as an effort to build a technological ecosystem that respects social justice and humanity. However, attention to the application of ethical principles in the development of artificial intelligence is still limited. In many countries, comprehensive guidelines to ensure that AI systems make fair, transparent and accountable decisions are still missing. Currently, ethical standards are often too general and non-binding, so companies only consider these principles as discourse without real implementation. The decision-making process is usually not transparent (black-box), which makes it difficult for users to understand the reasoning or logic behind the algorithms, resulting in decreased trust in the system. In addition, there is no clear structure of responsibility; when AI makes a mistake, there is often confusion over who is responsible between the developer, user, or AI provider, with no ability to determine who should be held accountable. This puts AI systems at risk of becoming “black boxes” that operate without adequate oversight and can be blamed when things go wrong.

This needs to be investigated further, especially at the local level, as more and more institutions and companies are starting to apply artificial intelligence in their decision-making processes. Without clear ethical guidelines, these systems can harm certain individuals or groups if the results of their decisions are biased or difficult to understand rationally. Therefore, an in-depth study of the ethical aspects of developing artificial intelligence systems that can automate decision-making is essential, so that their application remains in accordance with human values. This article aims to explore ethical issues in the development of AI systems that have the

capacity to make their own decisions. The focus of the discussion includes the moral responsibility of the developers, justice, openness, and accountability for the decisions made. Through this study, it is hoped that ethical principles that are relevant to the needs and context of Indonesian society in the face of advances in digital technology can be developed.

Research Methods

This study adopts a qualitative approach with a multiple case design, with the aim of deeply understanding the ethical interactions in the development of an artificial intelligence system that can make decisions autonomously. Data were collected through literature review, semi-structured interviews, and document analysis. The literature review included a review of recent sources related to the application of ethical principles in the field of AI technology, including the categories and tools used to translate ethics into practice, as indicated by the following (Morley et al. 2023) and (Becker et al. 2023). Interviews were conducted with 8 to 12 respondents consisting of technology developers, ethicists or social scientists, and policy makers or regulators. The focus of the questions revolved around how ethics are integrated, difficulties in applying principles of fairness, transparency, responsibility, and forms of accountability. In addition, document analysis was conducted on internal policies, algorithm audit results, and system design notes to examine the application of principles such as fairness and transparency in practice.

Data were analyzed using thematic analysis, with themes such as fairness, moral responsibility, transparency and accountability drawn from multiple data sources. The validity of the research results was maintained through triangulation between findings from literature reviews, interviews, and document analysis. In addition, a comparative study between the two research cases was used to look at similarities and differences in ethical implementation strategies between government agencies and private companies. This research framework refers to the model of the application of ethical principles in AI developed by (Morley et al. 2021), which emphasizes the need for concrete tools to implement high principles in the technology development cycle. This framework is supported by the model of “Ethics as a Service,” in which ethical responsibility is shared between an external ethics board and internal corporate developers. To ensure the validity and consistency of the data, this study applied a respondent-checking procedure, which meant verifying the findings with the informants, as well as documenting the entire process of data collection and analysis. Research ethics were observed through the application of the principles of informed consent, confidentiality of informants' identities, and omission of personal data. With this method, it is hoped that this research can provide a complete understanding of how the practical application of ethical principles in AI, the challenges faced by developers and stakeholders, and the opportunity to create a technology ethics framework that fits the values and conditions in Indonesia.

Results And Discussion

This research found that the operation of ethical principles in the development of artificial intelligence (AI) systems that make automated decisions faces systemic challenges. The two case studies analyzed, one in a government agency and one in the private sector, identified four main challenges: disharmony in internal audit standards, lack of clarity in audit coverage, weak organizational change management, and difficulty measuring the impact of audits on system performance.

An in-depth analysis of two case studies, one in government and one in the private sector, revealed four major systemic challenges in applying ethical principles to AI systems:

1. Disharmony of internal audit standards: Different organizations have different audit guidelines, so evaluations of fairness and moral responsibility are inconsistent and difficult to measure objectively. This variation in standards creates uncertainty about whether an AI system meets ethical expectations.
2. Vagueness of audit scope: Many parts of the AI development cycle-such as data sources, predictive models, and operational deployments-are not covered by ethical evaluation. This creates blind spots, where potential biases or systemic errors can go unnoticed, as also identified in audit studies in pharmaceuticals.
3. Weak change management: There is little ethics training for internal teams, and organizations rarely appoint change agents whose role is to ensure ethical values are implemented across the board. As a result, ethics do not become part of the organizational culture, but rather are treated temporarily and sporadically.
4. Difficulty measuring audit impact: Organizations struggle to assess the extent to which ethics audits contribute to improved system performance due to a lack of quantitative metrics and systematic reporting. This makes the evaluation of audit success only qualitative and subjective.

Table 1 Summary of Challenges

Barriers	Main Impact
Internal audit standards are not harmonized	Inconsistent evaluation of fairness; difficult to compare and improve practices
Audit scope is unclear	Blind spots appear in important parts; bias can go undetected
Weak change management	Ethics are not internalized; no change agent cares about applying principles
Difficult to measure audit impact	No quantitative data to assess audit effectiveness in improving AI

This explanation emphasizes that AI systems require not only ethical guidelines, but also strong and structured institutional procedures from audit design to impact evaluation so that the principles of fairness, transparency, and accountability can truly be implemented in a comprehensive and measurable manner.

This research shows that the use of differentially-privileged synthetic data and the integration of human-in-the-loop (HITL) mechanisms significantly improve fairness in AI systems, especially in models using COMPAS and UCI Adult datasets. The synthetic data is generated using privacy-preserving methods that preserve the statistical patterns and fairness characteristics of the original data, and tested through various fairness metrics. There was a marked improvement in indicators such as Parity Fairness, Equalized Odds, and Demographic Parity after the intervention.

Experiment by (Yuan and Wang 2025) Using synthetic differentially-privileged data shows that the fairness metrics on the original and synthetic datasets are highly correlated, proving that fairness is maintained after data intervention.

Furthermore, a study by (Giddens and Liu 2024) Using the SAFES approach-a combination of differential privacy and fairness-aware data transformation-confirms these findings by noting improved fairness in the COMPAS and UCI Adult datasets.

Table 1 Result Fairness Meter Before vs After Intervention

Dataset	Fairness Indicator	Before	After
COMPAS	Parity Fairness	0,59	0,72
COMPAS	Equalized Odds	0,65	0,83
UCI Adult	Demographic Parity	0,68	0,85

This method has not only proven to be effective on a research scale, but also has potential applications in the real world. The use of synthetic data allows organizations to protect individual privacy (through differential privacy), while improving the fairness of the system. The implementation of HITL ensures that there is human oversight ready to act when potential AI bias or decision errors arise.

As a concrete example, a company's internal team can train the model using synthetic data first, check fairness metrics, and then test the model on real data for validation. If any unfairness is found, rapid iteration using HITL allows for correction before the system goes into production.

1. The Ethics as a Service model is designed to bridge the gap between high-level AI ethics principles and technical operational practices. This approach resembles the structure of cloud computing combining the power of external guidance and internal flexibility. In essence, EaaS shares responsibility.
2. External ethics body (similar to a “platform”): sets ethical guidelines, audit standards, and conducts periodic external audits.
3. Internal company team (analogous to “infrastructure”): in charge of interpreting the principles into the local context, selecting and implementing fairness-aware tools, recording processes, and providing open ethical documentation.

This structure is supportive:

1. Multiple audits - internal audit by the team, plus external audit by the ethics body, thus minimizing the risk of incomprehensive ethics practices (ethics washing).
2. Continuous feedback loop - audit results are used to improve AI processes, design, and policies.
3. Clear division of roles - external maintains objective standards, internal ensures flexibility and adaptability to organizational context.

Thus, EaaS does not just make ethical principles a document, but serves as a structured and dynamic governance framework. This approach enables the application of ethics in a tangible, recorded and auditable manner, thus creating an AI system that is not only legal and innovative, but also responsible and human-centric.

Measuring the extent to which ethics audits have a tangible impact on the performance of AI systems has proven challenging. The absence of standard quantitative metrics, especially for aspects such as fairness, transparency, and accountability, makes it difficult for organizations to assess the effectiveness of their audits. Meanwhile, attributing impact is problematic because the complexity of the system and the many external factors that influence outcomes make it

difficult to identify the true contribution of audits to system improvement. Additionally,

The Challenge of Measuring the Impact of AI Ethics Audits	Brief Explanation
No standardized metrics	Difficult to compare audit benefits across organizations (wired.com)
Attribution difficulties	System complexity makes audit contributions difficult to isolate
Measurement bias	Choice and interpretation of metrics can distort audit results
Subjective judgment dependent	Audits without objective tools result in inconsistent evaluations

potential measurement biases in both metric selection and data interpretation can lead to misleading or inaccurate audit results.

Empirical studies also show that many AI audits have not been formally verified, or use only qualitative approaches without objective support tools, making success assessments heavily reliant on subjective human judgment. Some have suggested that AI audits should take a multimetric approach, combining quantitative data and qualitative narratives, and involving affected stakeholders in the process. However, to date, there are few concrete examples of standardized metrics being used consistently in combination with independent external audits.

The research confirms that to increase the value of ethics audits, organizations need to develop standardized quantitative metrics, transparent and bias-aware measurement methods, and systematic reporting systems-complete with independent audits and stakeholder engagement-so that audits are not merely rhetorical, but actually result in tangible and measurable improvements in system performance.

Discussion

This research confirms that while awareness of the importance of ethics in AI development has increased, its implementation in the field is still plagued by structural and procedural obstacles. Findings such as disharmonized audit standards, limited audit scope, and weak internal change management point to complex and systemic issues. Studies (Morley et al. 2021) support these results, highlighting that AI ethical principles documents are often ineffective in practice-too flexible to the point of “ethics washing”, or too rigid and decontextualized.

(Mökander and Floridi 2023) also introduced the concept of ethics-based auditing (EBA) as a holistic and sustainable governance framework. They suggested an ongoing auditing approach throughout the AI development cycle, with complete coverage to avoid blind spots. Their findings on barriers to standardization of audits, narrow coverage, and local resistance in national and global organizations also parallel the findings in Indonesia.

As an integrative solution, the Ethics as a Service (EaaS) model by (Morley et al. 2021) offers an adaptive and layered governance structure. This model separates functions: external ethics bodies set principles and conduct audits, while internal teams translate those principles into technical practices-with tools such as human-in-the-loop and fairness-aware tools-and document implementation transparently.

From a technical perspective, the experiment (Yuan and Wang 2025) and (Giddens and Liu 2024) shows that the use of synthetic data based on differential privacy, plus HITL, significantly improves fairness such as Parity Fairness and Equalized Odds on COMPAS and

UCI Adult datasets. This shows that technical methods are not only discussed on the theoretical stage, but also operational and efficient in real contexts.

Conclusions

This research aims to explore in depth how ethical principles are operationalized in the development of artificial intelligence (AI) systems that have the ability to make decisions automatically. Based on the analysis of literature, interviews with stakeholders, and review of policy and audit documents, it can be concluded that the process of applying ethics in AI is still partial and has not been thoroughly integrated into the technology development cycle. One of the key findings of this study is that many organizations face obstacles in translating ethical principles such as fairness, transparency, and accountability into technical and operational practices. The lack of uniform standards leads to variability in the conduct of ethics audits, while the scope of audits is often limited to documentation rather than ongoing AI systems. This creates potential blind spots in the moral and social oversight process. In addition, the lack of internal organizational awareness and capacity to build an ethical culture is also an obstacle. Ethics is not yet part of the daily decision-making process due to lack of training, absence of supporting structures such as ethics committees, and non-integration of ethical evaluation indicators in system performance measurement. Nonetheless, technical approaches such as the use of synthetic data and human involvement in decision-making have shown significant potential in improving algorithmic fairness metrics.

Governance models such as Ethics as a Service are a relevant solution in sharing ethical responsibilities between internal organizations and independent external institutions. This approach is considered capable of strengthening accountability and providing wider participation space for multidisciplinary actors. Therefore, the integration of ethical principles in AI systems needs to be understood not as a one-time process, but as a series of continuous cycles that require continuous reflection, evaluation, and adaptation. Thus, the conclusion of this study confirms that the application of ethics in AI system development requires institutional commitment, measurable practical instruments, and governance that is responsive to social and technological dynamics. The success of ethics integration depends not only on the good intentions of the developer, but also on the extent to which the organization is able to build structures, cultures, and oversight mechanisms that support ethics as an integral part of technology design.

Reference

- Becker, Sarah J, André T Nemat, Simon Lucas, René M Heinitz, Manfred Klevesath, and Jean Enno Charton. 2023. "A Code of Digital Ethics: Laying the Foundation for Digital Ethics in a Science and Technology Company." *AI & SOCIETY* 38 (6): 2629–39.
- Giddens, Spencer, and Fang Liu. 2024. "SAFES: Sequential Privacy and Fairness Enhancing Data Synthesis for Responsible AI." *ArXiv Preprint ArXiv:2411.09178*.
- Mökander, Jakob, and Luciano Floridi. 2023. "Operationalising AI Governance through Ethics-Based Auditing: An Industry Case Study." *AI and Ethics* 3 (2): 451–68.
- Morley, Jessica, Anat Elhalal, Francesca Garcia, Libby Kinsey, Jakob Mökander, and Luciano Floridi. 2021. "Ethics as a Service: A Pragmatic Operationalisation of AI Ethics." *Minds and Machines* 31 (2): 239–56.

Morley, Jessica, Libby Kinsey, Anat Elhalal, Francesca Garcia, Marta Ziosi, and Luciano Floridi. 2023. “Operationalising AI Ethics: Barriers, Enablers and next Steps.” *AI & SOCIETY*, 1–13.

Yuan, Chih-Cheng Rex, and Bow-Yaw Wang. 2025. “Quantitative Auditing of AI Fairness with Differentially Private Synthetic Data.” *ArXiv Preprint ArXiv:2504.21634*.