International Journal of Education, Vocational and Social Science
https://e-journal.citakonsulindo.or.id/index.php/IJEVSS
E-ISSN : 2963 4954
IJEVSS

Artificial Intelligence Algorithm Bias And Islamic Ethics According To Nurcholish Madjid: Formulating Inclusivity, Religious Humanism, And Substantive Justice As An Ethical Framework For Ai

Shodiqin
Universitas Paramadina, Indonesia

Corresponding Author: shodiqin@students.paramadina.ac.id

Abstract

The development of artificial intelligence (AI) has brought about significant transformations in various aspects of life, ranging from public services, education, and healthcare to the economic sector. However, various studies indicate that AI systems are not always neutral. The phenomenon of algorithmic bias demonstrates that AI can reproduce and even reinforce social injustice through discrimination based on race, gender, ethnicity, or specific social groups. Most studies on algorithmic bias are still dominated by Western ethical approaches that focus on the principles of fairness, accountability, transparency, and responsible AI, while contemporary Islamic ethical perspectives remain relatively limited. This study aims to analyze algorithmic bias from the perspective of Nurcholish Madjid's Islamic ethics, focusing on three main concepts: inclusivity, religious humanism, and substantive justice. This study is a literature review (library research) employing a philosophical and conceptual approach. Primary data were drawn from the works of Nurcholish Madjid, while secondary data were drawn from the literature on AI ethics, algorithmic bias, and technological justice. Data analysis was conducted using content analysis. The research findings indicate that Nurcholish Madjid's philosophy holds strong relevance in addressing the issue of algorithmic bias. The principle of inclusivity offers a critique of algorithmic discrimination and digital exclusion; religious humanism serves as the foundation for the development of human-centered AI that respects human dignity; while substantive justice provides a framework for evaluating the social impact of AI that goes beyond procedural justice. This research has resulted in the formulation of an AI ethics model based on Nurcholish Madjid's thought, built upon three main pillars: inclusivity, religious humanism, and substantive justice as a conceptual contribution to the development of Islamic AI ethics that is fair, inclusive, and humanistic.

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

Keywords: Algorithmic Bias, Artificial Intelligence, Islamic Ethics, Nurcholish Madjid, Religious Humanism.

INTRODUCTION

Artificial Intelligence (AI) has evolved from an experimental computational technology into a critical infrastructure that significantly influences various decision-making processes in society. Its applications are now widely adopted across sectors such as healthcare, education, economics, and public governance. Despite its numerous benefits, the advancement of AI has also introduced increasingly complex ethical challenges. One of the central issues in contemporary AI discourse is algorithmic bias, a condition in which algorithmic systems systematically produce outcomes that advantage or disadvantage specific groups (Barocas & Selbst, 2016; Mehrabi et al., 2021). Empirical evidence from cases such as COMPAS (Angwin et al., 2016), facial recognition systems (Buolamwini & Gebru, 2018), and Amazon's recruitment algorithm (Dastin, 2018) demonstrates that AI systems may generate biased decisions against specific groups based on race, gender, and other social identities. These cases highlight that algorithmic bias is not merely a technical problem, but also an ethical concern with profound implications for social justice (Floridi et al., 2018).

Although these approaches make important contributions, most of them originate from Western philosophical and ethical traditions. Studies on AI ethics from an Islamic perspective remain relatively limited, particularly those that connect the issue of algorithmic bias with the thought of contemporary Islamic intellectual figures. In fact, the Islamic intellectual tradition offers a rich source of values that can be employed to respond to ethical challenges arising from modern technological developments (Auda, 2008). One relevant figure in this context is Nurcholish Madjid (1939–2005). As one of Indonesia's most prominent Islamic thinkers, Nurcholish Madjid developed the idea of Islamic renewal that emphasizes a harmonious relationship between religion, humanity, and modernity. His thought is grounded in the principles of monotheism (tawhid), universal humanism, inclusivity, freedom, and social justice (Madjid, 1992, 1995).

Nevertheless, most of these approaches remain dominated by Western ethical traditions and tend to be procedurally oriented. As a result, the moral dimensions related to human dignity, diversity, and substantive social justice have not been fully accommodated. Contemporary AI ethics frameworks such as fairness, accountability, transparency, and explainability (FATE) have

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

contributed significantly, yet they are often criticized for being overly technical and insufficiently grounded in moral philosophy (Jobin et al., 2019; Binns, 2018). Therefore, a broader ethical foundation is required.

Three key concepts in Nurcholish Madjid's thought that are directly relevant to AI ethics are inclusivity, religious humanism, and substantive justice. These concepts emphasize recognition of human diversity, dignity, and justice beyond procedural formality. In contrast to dominant AI ethics approaches, this framework prioritizes human moral values and social responsibility in technological systems. In the context of algorithmic bias, such an approach enables evaluation not only of algorithmic performance but also of its broader social and ethical impacts on vulnerable groups and human dignity (Floridi et al., 2018; Kersten, 2013).

Although research on Artificial Intelligence Ethics (AI Ethics) has advanced significantly in recent years, several academic gaps remain. First, most contemporary studies on AI ethics focus on principles such as fairness, accountability, transparency, and explainability, which are largely rooted in Western ethical traditions (Jobin et al., 2019; Floridi et al., 2018). While these approaches make important contributions to mitigating algorithmic discrimination, they remain largely procedural and technical in orientation (Binns, 2018). Algorithmic bias itself has been widely documented as a systemic issue arising from data and model design, producing discriminatory outcomes in machine learning systems (Barocas & Selbst, 2016; Mehrabi et al., 2021).

Second, studies on AI ethics from an Islamic perspective have begun to emerge through approaches such as maqāṣid al-sharī'ah and ethical governance frameworks; however, relatively few of these works directly address algorithmic bias as a central issue. Third, scholarship on Nurcholish Madjid's thought has predominantly focused on themes of pluralism, democracy, and Islamic humanism, with limited engagement in contemporary technological ethics discourse (Bakti, 2005; Ichwan, 2012). Madjid's intellectual legacy emphasizes inclusive Islam, humanistic values, and contextual interpretation of religious principles in modern society (Hardiman, 2015; Kersten, 2013).

Based on these gaps, this study proposes a model of AI Ethics grounded in Nurcholish Madjid's intellectual framework by integrating the concepts of inclusivity, religious humanism, and substantive justice as a normative lens for analyzing algorithmic bias in artificial intelligence systems.

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

Tabel 1. State of the Art dan Research Gap

Research Domain	Research Focus	Main Findings	Study Limitations
Contemporary AI Ethics	Fairness, Accountability, Transparency, Explainability, Responsible AI	Provides a set of ethical principles designed to minimize algorithmic bias and enhance the accountability and transparency of artificial intelligence systems.	This approach tends to emphasize procedural and technical dimensions and remains limited in integrating Islamic ethical perspectives.
AI Ethics in an Islamic Perspective	Maqāṣid al-Sharī'ah, Maslahah, AI Governance, Moral Responsibility	Provides an Islamic normative foundation for the development and governance of artificial intelligence systems.	However, it has not specifically employed the thought of Nurcholish Madjid as an analytical framework for examining algorithmic bias.
Studies on Nurcholish Madjid's Thought	Pluralism, Democracy, Islamic Humanism, Modernization, and Social Justice	Demonstrates the relevance of Nurcholish Madjid's thought in addressing modern social and humanitarian issues.	However, it has rarely been connected to artificial intelligence, AI ethics, and algorithmic bias.
This Study	Algorithmic Bias in the Perspective of Nurcholish Madjid's Islamic Ethics	Formulates an AI ethics model based on inclusivity, religious humanism, and substantive	Offers a novel conceptual formulation that extends the

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

Research Domain	Research Focus	Main Findings	Study Limitations
		justice to address algorithmic bias.	discourse of Islamic AI ethics.

Based on the review of previous studies presented in Table 1, no research has been identified that systematically integrates the concepts of inclusivity, religious humanism, and substantive justice in Nurcholish Madjid's thought as a unified normative framework for analyzing algorithmic bias. Existing studies on AI ethics are largely dominated by principles such as fairness, accountability, transparency, and explainability (FATE), which are predominantly grounded in Western ethical and philosophical traditions (Jobin et al., 2019; Floridi et al., 2018). While these frameworks provide important foundations for addressing algorithmic discrimination, they tend to remain procedural, technical, and insufficiently rooted in broader moral and cultural worldviews (Binns, 2018).

In parallel, scholarship on algorithmic bias has extensively documented how machine learning systems may reproduce and amplify social inequalities through biased data, model design, and institutional structures (Barocas & Selbst, 2016; Mehrabi et al., 2021). However, these discussions rarely engage non Western ethical frameworks in a systematic manner. Similarly, although Islamic approaches to AI ethics have begun to emerge particularly through maqasid al shariah and maslahah based reasoning, these studies have not yet fully addressed algorithmic bias as a central analytical concern in a structured theoretical model.

Furthermore, studies on Nurcholish Madjid's intellectual tradition have primarily focused on themes such as pluralism, democracy, Islamic modernization, and humanism (Kersten, 2013; Bakti, 2005). His key ideas namely inclusivity, religious humanism, and substantive justice have not yet been systematically operationalized within the field of AI ethics or algorithmic governance. This reveals a clear conceptual and theoretical gap in the literature.

This gap indicates a significant opportunity for novelty in AI ethics scholarship, particularly in developing an Islamic ethical framework that is contextually responsive to contemporary technological challenges. Based on this gap, the present study addresses two main research questions namely how algorithmic bias in artificial intelligence can be understood from the perspective of Nurcholish Madjid's Islamic ethics and how the concepts of inclusivity, religious humanism, and substantive justice can be formulated into an Islamic AI Ethics Framework model.

This study aims to analyze algorithmic bias through the lens of Nurcholish Madjid's Islamic ethics while simultaneously developing an Islamic AI ethics model that is just, inclusive, and

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

humanistic, thereby contributing to a more context sensitive and normatively grounded discourse in AI ethics.

RESEARCH METHODOLOGY

This study employs a library research design with a philosophical conceptual approach and an interdisciplinary framework that connects Islamic studies, philosophy of technology, and artificial intelligence ethics. Library research is a method that relies on written materials as the primary source of data for analysis and interpretation without field data collection (Zed, 2004). This approach is suitable for studies that aim to construct theoretical and normative frameworks based on existing literature (Creswell & Poth, 2018).

The philosophical conceptual approach is used to critically examine, compare, and synthesize ideas across disciplines in order to build a coherent ethical framework. This approach enables the integration of Islamic ethical thought with contemporary debates in technology ethics and artificial intelligence governance (Krippendorff, 2018; Bowen, 2009).

The research data consist of primary and secondary sources. Primary data are derived from the works of Nurcholish Madjid, particularly Islam, Doktrin dan Peradaban, Islam, Kemodernan dan Keindonesiaan, Pintu Pintu Menuju Tuhan, and Kontekstualisasi Doktrin Islam dalam Sejarah. Secondary data are obtained from academic books, peer reviewed journal articles, reports from international organizations, and policy documents discussing algorithmic bias, artificial intelligence ethics, fairness, human centered artificial intelligence, artificial intelligence governance, and digital justice (Jobin et al., 2019; Floridi et al., 2018; Mehrabi et al., 2021).

Data collection is conducted through documentary analysis by selecting relevant literature based on its relevance to algorithmic bias, artificial intelligence ethics, and Nurcholish Madjid's intellectual framework. Documentary analysis is a systematic method for reviewing and interpreting documents to extract meaning and develop conceptual understanding (Bowen, 2009).

The unit of analysis focuses on three core concepts in Nurcholish Madjid's thought. First, inclusivity which includes non discrimination, data representation, and social plurality. Second, religious humanism which emphasizes human dignity, autonomy, and human oversight in technological systems. Third, substantive justice which focuses on social impact, protection of vulnerable groups, and equitable distribution of technological benefits (Kersten, 2013; Bakti, 2005).

Data analysis is conducted using qualitative content analysis through a structured multi

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

stage process. First, identification of algorithmic bias concepts in artificial intelligence ethics literature. Second, identification of ethical concepts in Nurcholish Madjid's thought. Third, thematic classification of both domains. Fourth, analysis of conceptual relationships between Islamic ethics and artificial intelligence ethics. Fifth, conceptual synthesis to construct an Islamic artificial intelligence ethics framework based on Nurcholish Madjid's perspective. Qualitative content analysis is a systematic technique for coding and interpreting textual data to produce conceptual insights (Krippendorff, 2018).

Research validity is ensured through source triangulation by comparing primary texts of Nurcholish Madjid, artificial intelligence ethics literature, and prior academic studies. Triangulation enhances the credibility and reliability of qualitative interpretation by cross verifying multiple sources of evidence (Denzin, 1978; Flick, 2018).

RESULTS AND DISCUSSION

ALGORITHMIC BIAS AS A CHALLENGE IN CONTEMPORARY AI ETHICS

The development of artificial intelligence (AI) has brought significant transformations across various sectors, including economics, education, healthcare, and public governance. Despite its strategic benefits, AI also presents serious ethical challenges, one of which is algorithmic bias. This issue demonstrates that AI systems do not always operate in a neutral manner, as they may reproduce social inequalities through automated decision-making processes that appear objective. Therefore, algorithmic bias cannot be understood merely as a technical issue, but must also be recognized as an ethical and social problem closely related to justice and fundamental human rights.

Understanding the Concept of Algorithmic Bias

In the literature on artificial intelligence ethics, algorithmic bias is understood as a systematic deviation that causes algorithmic decisions to fail to fully reflect principles of fairness and equality. Such bias may emerge at various stages of AI system development, including data collection, model training, variable selection, and system deployment within specific social contexts.

One of the primary sources of algorithmic bias originates from historical data used as the foundation for machine learning processes. When such data contain discrimination or social inequalities, algorithms tend to reproduce the same patterns in their outputs. In addition, the lack of representation of certain groups within datasets may result in AI systems performing more

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

accurately for majority groups than for minority groups. This condition indicates that AI does not generate bias independently, but rather often reinforces existing societal inequalities.

In this context, algorithmic bias demonstrates that technology is never entirely neutral. AI systems always operate within specific social structures and may influence the distribution of opportunities, rights, and access to resources. Therefore, the evaluation of AI systems cannot be limited to technical aspects alone, but must also incorporate broader social and ethical dimensions.

The COMPAS Case and Discrimination in the Criminal Justice System

One of the most widely cited cases in discussions of algorithmic bias is the use of COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) in the United States. This system is designed to predict the likelihood that a defendant will reoffend after being released, and its outputs are used as part of decision-making processes in criminal justice proceedings.

Initially, COMPAS was viewed as an effort to enhance objectivity within the legal system. Algorithmic systems were considered capable of reducing human judgment biases that often arise in decision-making processes. However, an investigation conducted by ProPublica in 2016 revealed that the system tended to assign higher risk scores to Black defendants compared to White defendants under similar conditions.

These findings sparked extensive debates on algorithmic fairness. Although COMPAS applies uniform procedures across all defendants, its outputs appear to produce disparate impacts across different racial groups. This case demonstrates that procedural equality does not necessarily lead to substantive justice.

From an ethical perspective, the COMPAS case illustrates that the use of AI in domains involving fundamental human rights requires strict oversight. When algorithmic decisions affect individual liberty, errors or biases may lead to severe social consequences.

From the perspective of Nurcholish Madjid's concept of substantive justice, the main issue in COMPAS is not merely algorithmic inaccuracy, but the social impact produced on specific racial groups. A system that appears procedurally neutral may still generate injustice when it reinforces existing discriminatory structures. Therefore, ethical evaluation of AI should not be limited to procedural fairness alone, but must also assess whether the system produces substantive justice in practice.

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

Facial Recognition and Data Representation Inequality

Another widely discussed case in the AI ethics literature is facial recognition technology. This technology is used across various sectors, including security, identity verification, and digital commercial services. Despite its potential benefits, research by Joy Buolamwini and Timnit Gebru shows that some facial recognition systems exhibit significantly higher error rates for women and individuals with darker skin tones compared to men with lighter skin tones.

This issue is primarily caused by unequal representation in training datasets. Most datasets used to train facial recognition algorithms are dominated by images of light-skinned men. As a result, the system develops higher accuracy in recognizing this group compared to underrepresented populations. This condition highlights the critical role of data quality and diversity in determining the fairness of AI systems. When datasets fail to adequately represent societal diversity, algorithms are likely to produce discriminatory outcomes against certain groups.

From the perspective of Nurcholish Madjid's concept of inclusivity, the problem of facial recognition is not only related to data imbalance, but also reflects a systemic failure to acknowledge human plurality. When algorithms perform better for majority groups than for minority groups, a form of digital exclusion emerges that contradicts the principle of respect for diversity.

Amazon Recruitment and the Reproduction of Gender Bias

Another illustrative case is Amazon's recruitment system, which demonstrates how algorithms can reproduce pre-existing social inequalities. Amazon developed an AI-based recruitment system to support hiring processes with the aim of improving recruitment efficiency.

However, the system was later found to assign lower scores to female applicants. The main cause was historical training data dominated by male hiring patterns. Consequently, the system learned that male-associated characteristics were more aligned with the company's hiring preferences than female-associated ones. This condition indicates that algorithms can reproduce gender stereotypes and reinforce existing social inequalities. It further demonstrates that AI systems cannot be separated from the social context in which data is generated, as embedded inequalities in data may be translated into automated decisions that appear objective and scientific.

From the perspective of Nurcholish Madjid's religious humanism, the core issue in Amazon's recruitment system lies in the tendency of AI to reduce human beings to historical data patterns. Individual assessment is no longer based on human potential and dignity in a holistic

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

sense, but rather on statistical patterns derived from past behavior. This condition risks eliminating the human dimension in decision-making processes.

Algorithmic Bias as a Contemporary AI Ethics Challenge

The aforementioned cases demonstrate that algorithmic bias is a multidimensional issue that cannot be adequately explained through a purely technical approach. The COMPAS case highlights issues of fairness in decision-making processes, facial recognition systems reveal problems of representation and digital exclusion, while Amazon's recruitment system illustrates how AI can reproduce historical inequalities embedded in training data.

These findings indicate that algorithmic bias is not merely a statistical error, but an ethical issue closely related to diversity, human dignity, and social justice. Although contemporary AI ethics has introduced various guiding principles such as fairness, accountability, transparency, and responsible AI, these approaches remain largely oriented toward procedural dimensions.

In this context, Nurcholish Madjid's thought offers a more substantive ethical perspective through three core principles: inclusivity, religious humanism, and substantive justice. These principles provide an important foundation for a more comprehensive analysis of algorithmic bias.

INCLUSIVITY AND CRITIQUE OF DIGITAL EXCLUSION

The development of artificial intelligence (AI) has created significant opportunities to improve human life through faster, more efficient, and more accurate services. However, these advancements also introduce new challenges in the form of digital exclusion arising from algorithmic bias. Digital exclusion does not always manifest as direct denial of access; rather, it often occurs through more subtle mechanisms such as unequal data representation, algorithmic discrimination, and automated decision-making that disadvantages certain groups. Because these decisions are produced by systems perceived as objective and scientific, the resulting forms of injustice are often difficult to detect.

This issue indicates that AI development requires not only technological sophistication but also a strong ethical foundation capable of ensuring respect for human diversity. In this context, Nurcholish Madjid's concept of inclusivity offers a relevant perspective for understanding and critiquing various forms of digital exclusion generated by algorithmic bias.

The concept of inclusivity in Nurcholish Madjid's thought aligns with the principle of fairness in AI ethics; however, it extends beyond procedural equality or statistical bias reduction. Inclusivity also emphasizes the recognition of social plurality, the protection of minority groups,

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

and respect for diversity as fundamental principles in the development of just technology.

Inclusivity in Nurcholish Madjid's Thought

Inclusivity is one of the central concepts in Nurcholish Madjid's intellectual framework. It is grounded in his view of universal humanism, which positions all human beings as creations of God endowed with equal dignity and value. According to Madjid, Islam does not promote exclusivism that claims a single group as the sole bearer of truth and legitimacy. Instead, Islam encourages openness to diversity and recognizes pluralism as part of divine will.

For Madjid, diversity in religion, ethnicity, culture, and social identity is an unavoidable reality of human existence. Such diversity is part of sunnatullah that must be accepted positively rather than used as a basis for discrimination. Therefore, an ideal society is one that manages differences through principles of equality, mutual respect, and justice.

This principle is strongly supported in the Qur'anic worldview:

يَا أَيُّهَا النَّاسُ إِنَّا خَلَقْنَاكُمْ مِنْ ذَكَرٍ وَأُنْثَى وَجَعَلْنَاكُمْ شُعُوبًا وَقَبَائِلَ لِتَعَارَفُوا إِنَّ أَكْرَمَكُمْ عِنْدَ اللَّهِ أَتْقَاكُمْ إِنَّ اللَّهَ عَلِيمٌ خَبِيرٌ

"O mankind! Indeed, We created you from a male and a female, and made you nations and tribes so that you may know one another." (Qur'an, Al-Ḥujurāt [49]: 13).

This verse indicates that plurality is part of a divine design intended to foster social relations based on mutual recognition (ta'āruf), rather than discrimination or domination of one group over another. Within the context of social ethics, inclusivity requires recognition of the existence and rights of every individual to receive fair treatment regardless of their social identity. This principle becomes particularly relevant when applied to the development of artificial intelligence technologies that have the capacity to significantly influence human life.

Algorithmic Bias as a Form of Digital Exclusion

From the perspective of inclusivity, algorithmic bias can be understood as a new form of social exclusion that occurs through digital technologies. When algorithms systematically produce outcomes that disadvantage certain groups, those groups experience reduced access to opportunities, resources, and social rights that should be equally available to them.

The cases of COMPAS, facial recognition systems, and Amazon's recruitment algorithm demonstrate that AI systems can generate digital exclusion even when they are not explicitly designed to discriminate. In the COMPAS case, Black defendants face a higher likelihood of receiving negative risk assessments from the system. In facial recognition systems, women and individuals with darker skin tones experience higher misidentification rates. Meanwhile, in

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

Amazon's recruitment system, female applicants receive fewer positive evaluations from automated selection processes.

These phenomena indicate that digital exclusion often originates from pre-existing societal inequalities that are subsequently reproduced through technological systems. Algorithms learn from historical data that reflect power relations, social stereotypes, and structural inequalities embedded in human society. As a result, AI may function as a mechanism that reinforces social injustice in a more systematic and less visible manner.

From Nurcholish Madjid's perspective, such conditions contradict the principle of inclusivity because they fail to recognize human dignity as equal across social groups. A system that systematically disadvantages certain groups reflects a failure to acknowledge diversity as a fundamental aspect of social reality that must be respected.

Data Representation and Recognition of Diversity

One of the fundamental issues contributing to algorithmic bias is the lack of representation of certain groups within datasets used to train AI systems. In many cases, available data predominantly represent majority groups rather than minority groups. As a result, algorithms tend to perform more accurately in recognizing characteristics of majority groups while producing higher error rates when dealing with underrepresented populations.

This issue demonstrates that data are never entirely neutral, as they reflect various social relations embedded within society. From the perspective of Nurcholish Madjid's inclusivity, diversity must be accommodated not only at the level of discourse but also within technological systems. Therefore, inclusive AI development requires more diverse data representation to ensure that algorithmic decisions do not disproportionately favor specific groups.

Inclusivity and the Protection of Minority Groups

The concept of inclusivity also demands the protection of minority and vulnerable groups that are most at risk of experiencing the negative impacts of algorithmic bias. Empirical studies indicate that underrepresented groups in datasets are more likely to face higher rates of erroneous algorithmic decisions compared to majority groups.

From Nurcholish Madjid's perspective, an inclusive society is not only one that recognizes diversity but also one that ensures equal participation and fair treatment for all its members. Therefore, the evaluation of AI systems should not be limited to accuracy metrics alone, but must

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

also consider their impacts on socially disadvantaged groups.

Inclusivity as a Foundation of AI Ethics

Based on the discussion above, it can be understood that the concept of inclusivity in Nurcholish Madjid's thought offers a profound critique of various forms of digital exclusion arising from algorithmic bias. Inclusivity does not merely signify openness to difference, but also requires the recognition of human diversity throughout the entire process of technological development and deployment.

In the context of AI, the principle of inclusivity can be operationalized through several measures. First, preventing algorithmic discrimination through continuous monitoring and evaluation of deployed systems. Second, ensuring diversity in datasets so that all social groups are adequately represented. Third, involving minority and vulnerable groups in the technology development process. Fourth, positioning respect for human plurality as a foundational principle in AI governance.

Thus, inclusivity functions not only as a critique of algorithmic bias, but also as a normative foundation for the development of more just and humane technologies. This perspective constitutes an important contribution of Nurcholish Madjid's thought to contemporary AI ethics discourse, while also opening pathways for the development of an Islamic AI ethics model grounded in respect for human diversity.

RELIGIOUS HUMANISM AND HUMAN-CENTERED AI

The development of artificial intelligence (AI) has brought fundamental changes to the relationship between humans and technology. Whereas technology previously functioned merely as a supportive tool that extended human capabilities, the emergence of AI has enabled it to play an increasingly significant role in decision-making processes. Algorithmic systems are now used to assess creditworthiness, recruit employees, determine healthcare priorities, manage public security, and even predict human behavior.

These developments raise a fundamental question regarding the position of human beings in the era of artificial intelligence. Do humans remain the central subject in technological development, or are they increasingly reduced to objects of data processed by algorithmic systems? This question is particularly important because many AI systems operate through data patterns, statistical categories, and behavioral predictions. In this context, the paradigm of human-centered AI emerges as a response that emphasizes the need to position humans as the

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

ultimate goal of technological development, rather than merely as data sources or objects of system optimization.

The principle of human-centered AI is closely aligned with the concept of religious humanism in Nurcholish Madjid's thought. Through this framework, human beings are understood as dignified entities endowed with freedom and moral responsibility that cannot be reduced solely to material or technical dimensions. In contrast to Western traditions of human-centered AI, Nurcholish Madjid's religious humanism introduces a transcendental dimension, whereby respect for human beings is grounded not only in individual rights but also in moral responsibility before God.

Religious Humanism in Nurcholish Madjid's Thought

Religious humanism is one of the central ideas in Nurcholish Madjid's intellectual framework. Unlike secular humanism, which places humans at the center of reality without reference to transcendence, religious humanism views human beings as dignified creatures due to their relationship with God.

According to Nurcholish Madjid, Islam is fundamentally a religion that deeply respects humanity. The presence of religion is not intended to eliminate human freedom, but rather to guide individuals in developing their potential within a framework of moral responsibility. Therefore, respect for human beings is an integral part of the doctrine of tawhid (monotheism).

In this view, human beings are not merely biological entities or social constructs, but also spiritual beings endowed with freedom of thought, moral capacity, and ethical responsibility. These dimensions collectively constitute human dignity, which must be respected in all aspects of life.

Madjid further emphasizes that scientific and technological progress must always be situated within a humanistic framework. Modernization should not result in human alienation from moral and spiritual values that form the foundation of existence. Instead, technology should function as a means to enhance human life and strengthen humanitarian values.

This perspective is highly relevant to the development of AI, as algorithmic systems often tend to perceive humans merely as data points that can be processed and predicted. Religious humanism serves as a reminder that human beings possess dimensions far more complex than information that can be fully captured by machines.

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

Human Reduction to Data in the AI Era

One of the main critiques of contemporary AI development is its tendency toward the reduction of humans into data (datafication). In modern digital systems, human activities are recorded, collected, and analyzed as data in order to generate predictions about individual and group behavior. While this approach can enhance system efficiency, it also carries the risk of reducing humans to statistical objects.

From the perspective of Nurcholish Madjid's religious humanism, human beings possess moral, spiritual, and freedom-based dimensions that cannot be fully represented through data alone. Therefore, the use of AI must avoid tendencies toward dehumanization that disregard the complexity of human experience.

Human Dignity as a Principle of AI Ethics

One of the key implications of religious humanism is the recognition of human dignity (human dignity). In Nurcholish Madjid's thought, human dignity originates from the status of human beings as God's creation, endowed with intrinsic value that cannot be measured solely in terms of productivity, social status, or economic capability.

This principle is strongly grounded in the Qur'anic worldview:

وَلَقَدْ كَرَّمْنَا بَنِي آدَمَ

"And indeed, We have honored the children of Adam." (Qur'an, Al-Isrā' [17]: 70).

This verse indicates that human dignity is an inherent gift bestowed upon all human beings regardless of race, religion, gender, or social background. Therefore, every social system is required to respect human dignity as a non-negotiable value.

In the context of AI, respect for human dignity requires that technology must not be used to degrade, exploit, or excessively control human beings. Algorithmic systems that produce discrimination, behavioral manipulation, or disproportionate surveillance are in contradiction with the principle of human dignity.

The principle of human dignity also demands the protection of individual privacy. Personal data is not merely an economic commodity to be traded, but part of human identity that must be respected and safeguarded. Therefore, AI development must take into account individuals' rights to privacy, freedom, and personal autonomy.

Thus, religious humanism offers a broader perspective compared to purely technical approaches that evaluate AI success solely in terms of accuracy and efficiency. Technological success must also be assessed based on its ability to preserve and enhance human dignity.

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

International Journal of Education, Vocational and Social Science
https://e-journal.citakonsulindo.or.id/index.php/IJEVSS
E-ISSN : 2963 4954
IJEVSS

Human Oversight in AI Systems

Another crucial aspect of the human-centered AI paradigm is the principle of human oversight, which refers to human control over algorithmic systems. This principle arises from the recognition that AI, regardless of its sophistication, does not possess moral consciousness or ethical responsibility in the same way humans do.

AI can generate recommendations, predictions, or decisions based on learned data patterns; however, it lacks the ability to understand moral values, empathy, and complex human contexts. Therefore, decisions with significant impacts on human life should not be fully delegated to algorithms.

From Nurcholish Madjid's perspective, human freedom is always accompanied by responsibility. Human beings are endowed with the capacity to think and develop technology, but they remain morally accountable for the consequences of its use. Accordingly, moral responsibility cannot be transferred to machines.

This perspective is consistent with a Prophetic tradition attributed to Prophet Muhammad (peace be upon him):

عَنِ ابْنِ عُمَرَ رَضِيَ اللَّهُ عَنْهُمَا قَالَ: سَمِعْتُ رَسُولَ اللَّهِ ﷺ يَقُولُ «كُلُّكُمْ رَاعٍ وَكُلُّكُمْ مَسْئُولٌ عَنْ رَعِيَّتِهِ، وَالإِمَامُ رَاعٍ وَمَسْئُولٌ عَنْ رَعِيَّتِهِ، وَالرَّجُلُ رَاعٍ فِي أَهْلِهِ وَهُوَ مَسْئُولٌ عَنْ رَعِيَّتِهِ، وَالْمَرْأَةُ رَاعِيَةٌ فِي بَيْتِ زَوْجِهَا وَمَسْئُولَةٌ عَنْ رَعِيَّتِهَا، وَالْخَادِمُ رَاعٍ فِي مَالِ سَيِّدِهِ وَمَسْئُولٌ عَنْ رَعِيَّتِهِ» (رواه البخاري ومسلم)

"Each of you is a leader, and each of you will be held accountable for those under your leadership. A leader is responsible for his people and will be questioned about them; a man is a leader within his household and will be held accountable for those under his care; a woman is a leader in her husband's home and will be held accountable for it; and a servant is a guardian of his master's wealth and will be questioned about it." Narrated by: Sahih al-Bukhari and Sahih Muslim

This hadith indicates that human beings remain moral subjects who are responsible for the decisions they make, including those involving the use of AI technology. In practical terms, the principle of human oversight requires mechanisms that enable humans to monitor, evaluate, and correct algorithmic decisions. Accordingly, AI functions as a supportive tool for humans rather than a replacement for human judgment in decisions that affect fundamental human rights.

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

Religious Humanism as the Foundation of Human-Centered AI

Based on the discussion above, it can be understood that the concept of religious humanism in Nurcholish Madjid's thought is highly relevant to the human-centered AI paradigm developed in contemporary technology ethics discourse. Both reject the view that places technology as the ultimate end and instead affirm that human beings must remain the central orientation of technological development.

In the context of AI, religious humanism can be translated into three core principles. First, technological development must be oriented toward human well-being (human-centered AI). Second, technology must respect human dignity and protect fundamental human rights. Third, human beings must retain control and responsibility over the use of AI (human oversight).

Thus, religious humanism does not merely function as a critique of digital dehumanization, but also offers a normative framework for the development of more civilized and ethically grounded AI systems. This perspective demonstrates that technological advancement does not have to be in conflict with religious and human values. On the contrary, technology can serve as a means to strengthen human dignity when developed on the basis of ethical principles that fully respect human values.

SUBSTANTIVE JUSTICE AS A CRITIQUE OF ALGORITHMIC FAIRNESS

In contemporary artificial intelligence (AI) ethics discourse, the issue of justice has become increasingly central in line with the growing use of algorithmic systems in various decision-making processes. Attention to this issue has intensified following the emergence of multiple cases of algorithmic bias, which demonstrate that AI systems can produce discriminatory outcomes against certain groups based on race, gender, ethnicity, religion, and other social characteristics. This condition has encouraged the development of fairness as one of the core principles in AI ethics.

Although fairness serves as an important foundation in efforts to reduce algorithmic discrimination, this approach has certain limitations. Most fairness models tend to emphasize statistical measures and procedural equality, such as equal treatment, equal opportunity, or balanced error distribution across groups. While such approaches help mitigate bias at a technical level, they do not always address more complex forms of injustice, particularly those related to the broader social impacts of technology deployment.

In this context, the concept of substantive justice developed by Nurcholish Madjid offers a broader perspective. Substantive justice does not merely ask whether procedures are applied equally, but also questions whether the outcomes and impacts generated truly reflect justice for

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

all social groups. Therefore, this concept can be used as a critique of fairness approaches that are overly focused on technical dimensions and insufficiently attentive to the surrounding social realities of technology use.

Justice in Nurcholish Madjid's Perspective

In Nurcholish Madjid's thought, justice is one of the fundamental values at the core of Islamic teachings. Justice is not merely understood as a legal rule or administrative mechanism, but as a moral principle that governs the relationship between humans and God, among individuals, and within society as a whole. According to Madjid, justice is a direct consequence of tawhid (monotheism). Awareness of the oneness of God leads to the recognition that all human beings possess equal standing as His creations. Therefore, there is no justification for discriminatory treatment or domination of one group over another.

This perspective is consistent with Qur'anic teachings that position justice as one of the primary objectives of social life:

يَا أَيُّهَا الَّذِينَ آمَنُوا كُونُوا قَوَّامِينَ لِلَّهِ شُهَدَاءَ بِالْقِسْطِ

"O you who believe! Stand firmly for justice for the sake of Allah..." (Qur'an, Al-Mā'idah [5]: 8).

This verse indicates that justice is not merely a moral choice, but an obligation that must be realized in all aspects of human life.

For Nurcholish Madjid, justice should not be confined to legal formalism. A system may appear procedurally fair while still producing injustice in practice. Therefore, the measure of justice must be based on the real impacts experienced by society, particularly by groups that are most vulnerable to injustice.

This perspective forms the foundation of the concept of substantive justice. Substantive justice emphasizes that a policy, system, or technology must be evaluated based on its social consequences rather than merely on the procedures through which it is implemented.

Limitations of Fairness in AI Ethics

In contemporary AI discourse, fairness is one of the most widely adopted principles for addressing algorithmic bias. Various mathematical models have been developed to assess whether an algorithm treats different groups equitably.

Common approaches include:

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

a. Demographic parity, which ensures that algorithmic outcomes are proportionally distributed across groups.
b. Equal opportunity, which ensures that individuals who meet certain criteria have equal chances of receiving positive outcomes.
c. Equalized odds, which balances error rates across different social groups.

These approaches make important contributions to identifying and mitigating algorithmic bias. However, many scholars argue that fairness often addresses issues only at the statistical level without engaging with broader structural inequalities.

For instance, a recruitment system may be considered fair because it provides equal opportunities to all applicants. However, if certain groups already experience lower access to education due to structural inequality, procedural equality does not necessarily result in substantive justice.

Similar issues can be found in digital credit systems, AI-based healthcare services, and predictive policing systems. Algorithms may operate under uniform procedures for all individuals, yet the resulting impacts can differ significantly across groups.

From the perspective of substantive justice, this condition indicates that justice cannot be measured solely through procedural equality. More importantly, it must assess whether technology reduces or instead reinforces existing inequalities.

Evaluating Social Impact as a Measure of Justice

One of the key contributions of the concept of substantive justice is its emphasis on social impact evaluation. From this perspective, the success of an AI system is not only measured by technical accuracy or efficiency, but also by its social consequences.

This approach is particularly important because many AI systems are deployed in domains that directly affect human lives, such as education, healthcare, employment, financial services, and law enforcement. Algorithmic decisions in these areas can determine individuals' access to social and economic opportunities.

The COMPAS case provides a clear example of the importance of social impact evaluation. Although the system was designed to improve objectivity in criminal justice, its outcomes disproportionately affected certain racial groups. From the perspective of substantive justice, such a system cannot be considered fair merely because it applies uniform procedures to all individuals.

Nurcholish Madjid emphasizes that justice must be realized in the form of tangible social

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

benefit (maslahah). Therefore, the evaluation of AI systems must consider whether technology provides equitable benefits to society or instead amplifies existing inequalities.

Thus, social impact evaluation becomes a crucial element in developing a more comprehensive and human-oriented AI ethics framework.

Protection of Vulnerable Groups

Another essential dimension of substantive justice is the protection of vulnerable groups. In many societies, certain groups occupy weaker social, economic, or political positions compared to others. These groups are often at higher risk of experiencing negative impacts from biased algorithmic systems.

Minority groups, women, low-income communities, persons with disabilities, and populations underrepresented in datasets are among those most vulnerable to algorithmic bias. They not only face a higher likelihood of erroneous decisions but also often have limited capacity to contest or correct such outcomes.

In Nurcholish Madjid's thought, solidarity with the (weak) constitutes part of moral responsibility in Islam. Justice does not mean treating everyone identically, but ensuring that those in vulnerable positions receive adequate protection so that they can enjoy equal rights.

This principle aligns with the Qur'anic injunction:

إِنَّ اللَّهَ يَأْمُرُ بِالْعَدْلِ وَالإِحْسَانِ

"Indeed, Allah commands justice, excellence..." (Qur'an, An-Naḥl [16]: 90).

In the context of AI, the protection of vulnerable groups can be operationalized through algorithmic bias audits, increased system transparency, the involvement of affected groups in technology development processes, and the provision of appeal mechanisms against automated decisions.

Accordingly, substantive justice requires special attention to groups that are at the highest risk of marginalization due to the deployment of technological systems.

DISTRIBUTION OF TECHNOLOGICAL BENEFITS AND SOCIAL JUSTICE

The final aspect of substantive justice concerns the distribution of technological benefits. The development of AI generates significant economic and social advantages. However, these benefits are not always distributed equally across all layers of society.

In many cases, the greatest advantages of digital transformation are enjoyed by large technology companies and groups with access to digital resources. Conversely, groups with limited

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

access to education, technology, and digital infrastructure often lag behind and receive only a small portion of the benefits generated.

Nurcholish Madjid argues that modernity should be directed toward creating collective welfare rather than widening social inequality. Technological advancement that benefits only certain groups is contrary to the principles of social justice embedded in Islamic ethics.

In the context of AI, the distribution of technological benefits becomes an important measure of whether a system truly contributes to public welfare (maslahah). Therefore, AI development must be accompanied by efforts to expand technological access, improve digital literacy, and ensure that the benefits of innovation are accessible to all segments of society.

SUBSTANTIVE JUSTICE AS A FRAMEWORK FOR AI ETHICS

Based on the discussion above, it can be understood that the concept of substantive justice in Nurcholish Madjid's thought offers a significant critique of the fairness approach that dominates contemporary AI ethics. While fairness primarily focuses on procedural and statistical equality, substantive justice emphasizes the importance of considering the social impacts produced by technology.

In the context of AI, substantive justice can be operationalized through three main principles:

a. Social impact evaluation, which assesses AI based on its real consequences for human life.
b. Protection of vulnerable groups, which ensures special attention to those most at risk of discrimination and marginalization.
c. Distribution of technological benefits, which ensures that the outcomes of AI development are more equitably accessible to all members of society.

Through these three principles, substantive justice functions not only as a critique of algorithmic bias but also as a normative foundation for the development of more human-centered and socially just technologies. Thus, Nurcholish Madjid's thought makes a significant contribution to expanding the scope of AI ethics from a purely technical concern to a broader concern for human welfare (maslahah).

Substantive justice extends the concept of algorithmic fairness by positioning social impact as the primary criterion for evaluating technological justice.

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

FORMULATION OF AN AI ETHICS MODEL BASED ON NURCHOLISH MADJID'S THOUGHT

The preceding discussion demonstrates that the problem of algorithmic bias in artificial intelligence cannot be understood solely as a technical issue related to data quality, algorithm design, or system accuracy. Algorithmic bias is fundamentally an ethical and social problem concerning how technology perceives human beings, manages diversity, and distributes benefits and risks within society. Therefore, efforts to address algorithmic bias require an ethical framework that not only focuses on procedural aspects but also responds to deeper human concerns.

Based on the analysis of Nurcholish Madjid's thought and various contemporary AI ethics issues, this study formulates an AI ethics model grounded in his intellectual framework. This model represents a conceptual synthesis that integrates Islamic ethical values with the challenges of artificial intelligence development. Unlike most AI ethics models derived from Western philosophical traditions, the proposed model is grounded in the principles of tawhid and universal humanism as its normative foundation.

From this foundation emerge three core principles that form the pillars of the Nurcholish Madjid-based AI Ethics framework, namely inclusivity, religious humanism, and substantive justice. These three principles are interrelated and collectively form an ethical structure aimed at achieving the development of AI that is fair, inclusive, and human-centered.

Tawhid as the Foundation of AI Ethics

In Nurcholish Madjid's thought, tawhid (the oneness of God) constitutes the core principle of Islam, encompassing not only theological dimensions but also social and ethical implications. Tawhid affirms the unity of God while simultaneously emphasizing the unity of humankind as His creation. This consciousness leads to the recognition that all human beings possess equal dignity and must not be treated discriminatorily based on social, economic, or cultural identities.

The principle of tawhid also implies that all human activities, including the development of science and technology, must be directed toward achieving maslahah (public good) and preventing harm. Technology should not be regarded as an ultimate goal, but rather as an instrument that supports human responsibility as vicegerents (khalifah) on earth.

Within the context of AI, tawhid serves as a moral foundation guiding technological development toward human-centered values. AI must not be used to create domination, exploitation, or injustice against specific groups. Instead, it should be directed toward expanding social benefits and improving human well-being. Thus, tawhid functions as a normative source

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

that connects technological ethics with human moral responsibility toward both fellow human beings and God.

Universal Humanism as a Moral Orientation

From the principle of tawhid emerges the concept of universal humanism, which represents a key characteristic of Nurcholish Madjid's intellectual framework. Universal humanism affirms that every human being possesses inherent value and dignity that must be respected regardless of religion, race, ethnicity, gender, or social status.

For Madjid, human diversity is not a threat but a divinely ordained reality that must be accepted. Therefore, a civilized society is one that manages differences through principles of equality, respect, and justice.

In the context of artificial intelligence, universal humanism serves as a moral orientation ensuring that technology is developed for all humanity, not for the benefit of specific groups alone. AI development must consider its impact on diverse social groups, particularly those who are historically underrepresented or have limited access to digital spaces.

Universal humanism also provides a normative basis for rejecting algorithmic discrimination arising from biased data or system design. Accordingly, all AI development processes must be oriented toward respecting human diversity and dignity in a universal manner.

Inclusivity as the First Pillar

The first principle derived from universal humanism is inclusivity. As discussed previously, inclusivity emphasizes the recognition of human diversity and the rejection of all forms of social and digital exclusion.

In the context of AI, inclusivity functions as a principle ensuring that technology does not produce discrimination against specific groups. This principle is operationalized through several key aspects. First, the prevention of algorithmic discrimination based on race, gender, religion, ethnicity, or other social identities. Second, data diversity to ensure that AI systems accurately represent social reality. Third, the representation of minority groups in technology development and evaluation processes. Fourth, stakeholder participation in AI governance.

Through inclusivity, AI systems are not only evaluated based on predictive accuracy but also on their ability to respect and reflect human diversity.

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

Religious Humanism as the Second Pillar

The second pillar of Nurcholish Madjid's AI ethics model is religious humanism. This principle positions human beings as the central orientation of technological development and rejects any form of dehumanization resulting from AI applications.

Religious humanism is grounded in the belief that humans possess dignity that cannot be reduced to data, numerical values, or statistical patterns. Human beings are moral agents endowed with freedom, responsibility, and a spiritual dimension that cannot be fully captured by algorithmic systems.

In the context of AI, religious humanism is manifested through three main principles. First, human-centered AI, which emphasizes technological development oriented toward human well-being. Second, human dignity, which includes the protection of privacy, freedom, and individual autonomy. Third, human oversight, which ensures that humans retain control over AI systems and remain responsible for their outcomes.

Through this principle, technology is positioned as a tool that serves humans rather than a force that controls them.

Substantive Justice as the Third Pillar

The third pillar is substantive justice. This principle offers a critique of the dominant fairness approach in contemporary AI ethics. While fairness primarily focuses on procedural and statistical equality, substantive justice emphasizes the importance of assessing the real-world impacts produced by technology on human life.

Substantive justice is operationalized through three main components. First, social impact assessment in the use of AI systems. Second, protection of vulnerable groups that are at higher risk of discrimination and marginalization due to algorithmic systems. Third, equitable distribution of technological benefits so that the advancement of AI can be experienced by all segments of society.

Through this approach, an AI system cannot be considered fair merely because it applies uniform procedures to all individuals. Instead, it must also demonstrate its capacity to produce more equitable outcomes and reduce existing social inequalities.

Nurcholish Madjid-Based AI Ethics Model

Based on the analysis of inclusivity, religious humanism, and substantive justice in Nurcholish Madjid's thought, this study formulates an AI Ethics model grounded in tawhid and

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

International Journal of Education, Vocational and Social Science
https://e-journal.citakonsulindo.or.id/index.php/IJEVSS
E-ISSN : 2963 4954
IJEVSS

universal humanism as its normative foundation. These three concepts are interconnected in forming an ethical framework that is not only oriented toward procedural fairness, but also toward respect for human dignity, recognition of diversity, and the creation of social justice. The formulation of this model is presented in Figure 1.



Figure 1. Conceptual Framework of Islamic AI Ethics Based on the Thought of Nurcholish Madjid.

The proposed model indicates that tawhid serves as an ethical foundation that generates awareness of universal humanism as a moral orientation in the development of artificial intelligence. This awareness is operationalized through three core pillars, namely inclusivity,

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

religious humanism, and substantive justice. These three pillars function as a normative framework for preventing algorithmic discrimination, respecting human dignity, and realizing social justice in the use of technology. Thus, the AI Ethics model based on Nurcholish Madjid's thought is directed toward the development of AI systems that are fair, inclusive, and human-centered.

Theoretical Contribution of the Study

The main finding of this study is the formulation of an Islamic AI Ethics model grounded in Nurcholish Madjid's intellectual framework. This model contributes theoretically in two main aspects.

First, it expands the scope of AI ethics studies by introducing a contemporary Islamic perspective that has been underrepresented in discussions on algorithmic bias. Second, it offers a synthesis between Islamic ethical values and modern technological development, resulting in an ethical framework that addresses not only technical issues but also broader concerns of humanity and social justice.

Accordingly, Nurcholish Madjid's thought is not only relevant as a critical instrument for examining algorithmic bias, but also serves as a normative foundation for developing AI systems that are more just, inclusive, and human-centered. This model represents the main finding of the study as well as a conceptual contribution to the development of Islamic AI ethics in the era of artificial intelligence.

CONCLUSION

This study demonstrates that algorithmic bias in artificial intelligence cannot be understood merely as a technical issue related to data quality or system design, but also as an ethical and social problem that affects justice, diversity, and human dignity. Through the lens of Nurcholish Madjid's Islamic ethics, algorithmic bias can be critically examined through three core concepts, namely inclusivity, religious humanism, and substantive justice. Inclusivity functions as a critique of algorithmic discrimination and digital exclusion, religious humanism provides a foundation for developing human-centered AI that respects human dignity, while substantive justice extends the concept of fairness by emphasizing social impact evaluation, protection of vulnerable groups, and more equitable distribution of technological benefits.

This study produces a conceptual formulation of an AI Ethics model based on Nurcholish Madjid's thought, grounded in the principles of tawhid and universal humanism and

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

operationalized through three main pillars: inclusivity, religious humanism, and substantive justice. This model offers an alternative perspective to prevailing AI ethics approaches that are largely dominated by procedural fairness paradigms, by introducing a stronger orientation toward human dignity, recognition of diversity, and substantive social justice.

Theoretically, this research contributes to the development of AI ethics studies from an Islamic perspective by extending the relevance of Nurcholish Madjid's thought into contemporary digital technology discourse. Furthermore, it demonstrates that Islamic values are not only relevant as a critical tool for examining algorithmic bias, but can also serve as a normative foundation for formulating AI ethics models that are just, inclusive, and human-centered. Therefore, the proposed model may serve as a conceptual framework for the development, evaluation, and governance of artificial intelligence that is more oriented toward human welfare in the digital era.

This study is conceptual in nature and does not empirically test the proposed framework on specific AI systems. Future research is encouraged to further develop this model through empirical studies and algorithmic auditing approaches.

REFERENCES

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*.

Auda, J. (2008). *Maqasid al-shariah as philosophy of Islamic law: A systems approach*. International Institute of Islamic Thought.

Bakti, A. F. (2005). Islam and modernity: Nurcholish Madjid's interpretation of civil society, pluralism, secularization, and democracy. *Asian Journal of Social Science, 33*(3), 486–510.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104(3), 671–732.

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. In Proceedings of the 2018 Conference on Fairness, Accountability and Transparency.

Bowen, G. A. (2009). Document analysis as a qualitative research method. *Qualitative Research Journal, 9*(2), 27–40.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of Machine Learning Research*, 81, 1–15.

Creswell, J. W., & Poth, C. N. (2018). *Qualitative inquiry and research design: Choosing among five approaches*. Sage Publications.

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

Dastin, J. (2018, October 10). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*.

Denzin, N. K. (1978). *The research act: A theoretical introduction to sociological methods*. McGraw Hill.

Flick, U. (2018). *An introduction to qualitative research*. Sage Publications.

Floridi, L., Cowls, J., Beltrametti, M., et al. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*, 689–707.

Hardiman, F. B. (2015). Nurcholish Madjid: Islam and pluralistic democracy in Indonesia. *Verbum SVD*, 56(2), 180–194.

Ichwan, M. N. (2012). The making of a Pancasila state: Political debates on secularism, Islam and the state in Indonesia. UIN Sunan Kalijaga.

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

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

Kersten, C. (2013). Cosmopolitan Muslim intellectuals and the mediation of cultural Islam in Indonesia. In *East by Mid-East: Studies in cultural, historical, and political Islam*.

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage Publications.

Madjid, N. (1992). *Islam, doktrin dan peradaban*. Paramadina.

Madjid, N. (1995). *Islam, kemodernan dan keindonesiaan*. Paramadina.

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.

Rambe, K. A., & Fatimah, S. (2026). Nurcholish Madjid's study of thought for the civilization of science. *Journal of Social Science Education, Art and Culture*.

Zed, M. (2004). *Metode penelitian kepustakaan*. Yayasan Obor Indonesia.

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License