PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 22, No. 2 September 2026 | DOI: 10.33480/pilar.v22i2.8531

UNCOVERING DIGITAL DISCRIMINATION: ALGORITHMIIC BIAS AND DIVINE VALUE RESPONSES IN THE BIG DATA ERA

Alisha Zulfa Salsabila*; Talitha Falakh Tasya; Najwa Amalia Khoir; Zaskiyah Sofarina; Muhamad Masrur

Data Science Study Program
UIN K.H. Abdurrahman Wahid, Pekalongan, Indonesian
https://uingusdur.ac.id/
alisha.zulfa.salsabila25010@mhs.uingusdur.ac.id*, talitha.falakh.tasya25016@mhs.uingusdur.ac.id, najwa.amalia.khoir25009@mhs.uingusdur.ac.id, zaskiyah.sofarina25013@mhs.uingusdur.ac.id, m.masrur@uingusdur.ac.id
(*) Corresponding Author



The creation is distributed under the Creative Commons Attribution-NonCommercial 4.0 International License.

Abstract*—AI systems now decide what content people see online, at scale. The engineers building them generally aim for fairness. The systems still discriminate against specific groups anyway. This study measures that bias in content moderation systems, focusing on religious, racial, and gender identity, then reads the results through Islamic ethics. We used a quantitative-descriptive approach and computational content analysis on the Jigsaw Unintended Bias in Toxicity Classification dataset 90,902 entries after balancing classes via proportional undersampling and measured bias with chi-square tests, independent t-tests, and fairness metrics including Demographic Parity Difference and Disparate Impact Ratio. The headline result: comments mentioning Muslim identity got approved 21.1 percentage points less often than comments mentioning Christian identity, 66.3% versus 87.4% (chi-square = 184.98, p < 0.001; Disparate Impact Ratio = 0.758, below the 0.8 line normally used to flag adverse impact). Three things in the training data drive this: annotation bias, historical context bias, and co-occurrence bias. We could check the third one directly the word "terrorist" shows up alongside Muslim-identity comments 13.3% of the time, against 2.5% for Christian-identity comments. We then read these results against four Islamic principles tabayyun (verification), 'adl (justice), musawah (equality), and amanah (accountability) within the Maqashid al-Shariah framework, which we define here as normative principles drawn from the Qur'an and Islamic jurisprudence governing just conduct, equality, and accountability. The paper's contribution is an Islamic AI Governance Framework that turns these principles into actual mechanisms: bias auditing, annotator diversification, algorithmic transparency, inclusive data governance.*

Keywords: *algorithmic bias, content moderation, digital discrimination, fair artificial intelligence, 0maqashid Al-Shariah.*

Abstrak—*Algoritma AI sekarang menentukan informasi apa yang kita lihat, dalam skala besar. Insinyur yang membangunnya umumnya peduli soal keadilan. Sistemnya tetap saja diskriminatif terhadap kelompok tertentu. Penelitian ini mengukur bias itu dalam sistem moderasi konten, dengan fokus pada identitas agama, ras, dan gender, lalu membaca hasilnya lewat etika Islam. Kami pakai pendekatan kuantitatif-deskriptif dan analisis konten komputasional pada dataset Jigsaw Unintended Bias in Toxicity Classification 90.902 entri setelah penyeimbangan kelas lewat proportional undersampling dan mengukur bias dengan uji chi-square, independent t-test, serta fairness metrics termasuk Demographic Parity Difference dan Disparate Impact Ratio. Temuan utamanya: komentar bermuatan identitas Muslim disetujui 21,1 poin persentase lebih jarang dibanding komentar bermuatan identitas Kristen, 66,3% berbanding 87,4% (chi-square = 184,98, p < 0,001; Disparate Impact Ratio = 0,758, di bawah ambang 0,8 yang biasa dipakai untuk menandai adverse impact). Tiga hal dalam data pelatihan mendorong ini: bias anotasi, bias konteks historis, dan bias ko-okurens. Yang terakhir bisa kami cek langsung kata "terrorist" muncul berdampingan dengan komentar beridentitas Muslim sebanyak 13,3%, melawan 2,5% pada komentar beridentitas Kristen. Hasil ini lalu kami baca lewat empat prinsip Islam tabayyun (verifikasi), 'adl (keadilan), musawah (kesetaraan), dan amanah (akuntabilitas) dalam kerangka Maqashid al-Syariah, yang kami definisikan sebagai prinsip normatif dari Al-Qur'an dan yurisprudensi*

Islam yang mengatur keadilan, kesetaraan, dan akuntabilitas. Kontribusi penelitian ini adalah Islamic AI Governance Framework yang mengubah prinsip-prinsip tadi jadi mekanisme nyata: audit bias, diversifikasi anotator, transparansi algoritma, tata kelola data inklusif.

***Kata Kunci**: bias algoritmik, moderasi konten, diskriminasi digital, kecerdasan buatan berkeadilan, Maqashid Al-Syariah.*

INTRODUCTION

In less than two decades, artificial intelligence (*Artificial Intelligence*/AI) systems and big data have worked their way into nearly every domain of daily life: job recruitment, credit assessment, law enforcement, healthcare recommendations. Algorithms no longer merely process information they filter, rank, and frame it, shaping what the public sees and considers credible (He, 2024). While these technologies offer substantial improvements in efficiency and decision-making, they also introduce significant ethical challenges, particularly the issue of algorithmic bias.

Algorithmic bias is the systematic tendency of a computational system to produce discriminatory outcomes against particular groups (Ferrara, 2024). The documented cases are not marginal. The COMPAS system used in U.S. courts was twice as likely to flag Black defendants as high-risk. Facial recognition tools from major tech companies failed to identify dark-skinned women at error rates reaching 34.7%. Resume-screening algorithms discriminated against women (Raquib, Channa, Zubair, & Qadir, 2022). These examples demonstrate that algorithmic bias is not an isolated technical issue but a broader social and ethical challenge.

Within digital content moderation, algorithmic bias manifests in a more subtle yet equally consequential manner. The Jigsaw Unintended Bias in Toxicity Classification dataset over 174,000 online comments collected from 2015 to 2017 shows that comments referencing Muslim identity consistently receive lower approval rates than those about other religious groups, Christians in particular. This is not a random statistical fluctuation. It reflects historical context bias, annotation bias, and co-occurrence bias that the model absorbed from its training data.

The presence of algorithmic bias in the religious domain carries profound epistemological implications. When algorithms function as epistemic gatekeepers determining which content is deemed relevant, credible, and worth seeing epistemic authority shifts away from scholars, commentators, and academic institutions toward computational systems that operate with no ethical awareness and no normative framework. Furthermore, algorithmically driven filter bubbles and echo chambers reinforce ideological polarization by limiting users' exposure to diverse perspectives, thereby increasing the risk of misunderstanding and discrimination in digital religious discourse.

Although research on AI ethics and Islamic values has expanded considerably in recent years, an important gap remains. Most existing work takes a secular approach to algorithmic bias without integrating a divine-value perspective. Studies that do engage a religious lens are typically limited to one tradition and remain normative-textual rather than empirically grounded. None, to our knowledge, has systematically analyzed algorithmic bias in real datasets and then responded to those findings through a cross-traditional divine-value framework (Yaakob, Abdullah, & Mohammad Masruh, 2025). This gap highlights the need for an interdisciplinary approach that bridges empirical data science with Islamic ethical thought to develop a more comprehensive framework for responsible AI governance.

Accordingly, this study aims to identify and measure the primary forms of algorithmic bias related to religious, racial, and gender identities through computational analysis of the Jigsaw Unintended Bias in Toxicity Classification dataset. Furthermore, this study explores and synthesizes Islamic divine values, particularly the concepts of *tabayyun* (QS. al-Hujurat: 6), *ta'aruf* (QS. al-Hujurat: 13), *'adl* (QS. al-Nahl: 90), and the *Maqashid al-Shariah* framework, as normative foundations for evaluating the ethical implications of algorithmic bias in digital environments (Binti Abu Bakar, Binti Hussin, Bin Ahmad, & Bin Rozaini, 2025). Finally, the study proposes an integrative framework that combines Qur'anic divine values with algorithmic fairness principles to support the development of AI governance that is dignified, inclusive, asccountable, and ethically grounded, thereby contributing to the broader discourse on value-based artificial intelligence ethics.

MATERIALS AND METHODS

Research Type and Approach

This study runs a quantitative-descriptive approach together with computational content analysis. The quantitative side measures and classifies patterns of algorithmic bias in structured numerical data. The descriptive side documents the discrimination that turns up in AI-based content moderatio systems (Abdul Aziz & Zulkepli, 2025).

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 22, No. 2 September 2026 | DOI: 10.33480/pilar.v22i2.8531

It sits within critical positivism a paradigm that treats social reality as measurable but doesn't pretend data analysis is value-free. That's the point: this isn't just a technical bias count. The measurements get read against an Islamic normative framework, *Maqashid al-Shariah.* The data themselves are secondary, drawn from a publicly available dataset.

Data Collection Technique

The data collection process was conducted systematically to ensure the reliability and reproducibility of the research. First, the Jigsaw Unintended Bias in Toxicity Classification dataset was selected because it is a publicly available benchmark dataset specifically designed to investigate unintended bias in machine learning models for content moderation. The dataset was downloaded from the official Kaggle repository (https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/data) after complying with the dataset's terms of use. Subsequently, data integrity was verified by examining the completeness and consistency of the CSV files using the Python pandas library to identify potential missing values, formatting inconsistencies, or corrupted records. To support research transparency and reproducibility, all relevant metadata, including the dataset version, release information, licensing terms, and original source, were documented throughout the data collection process.

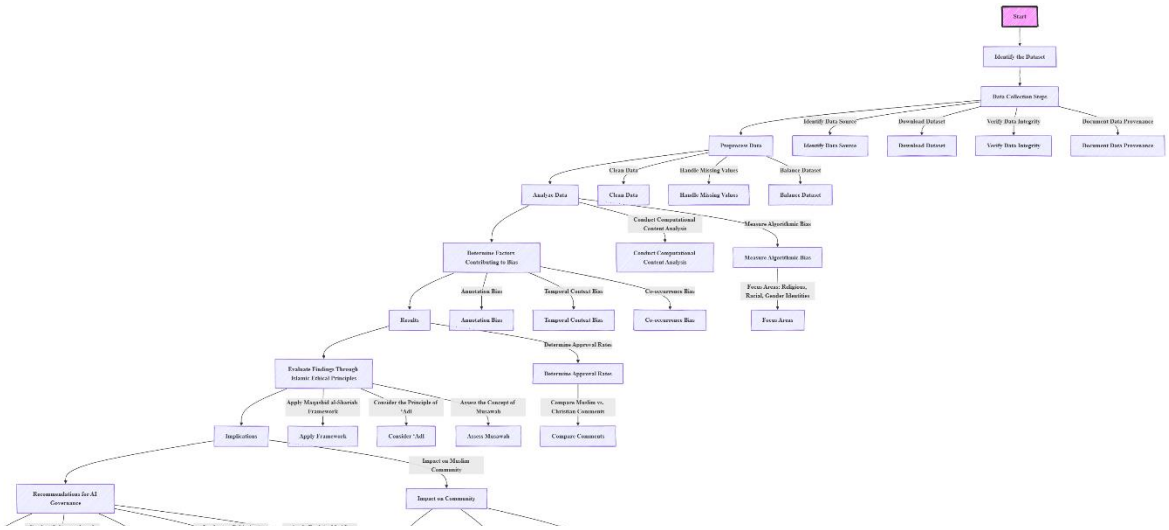
The selection of this dataset was based on several considerations. First, it has been widely adopted in studies on algorithmic fairness and AI ethics, making it a reliable benchmark for bias analysis. Second, the dataset explicitly contains identity-related attributes, including religious identities such as Muslim, Christian, Hindu, Buddhist, Jewish, and Atheist, which are directly aligned with the objectives of this study. Finally, with 174,716 annotated comments, the dataset provides sufficient data volume to support statistically robust and representative analyses of algorithmic bias across multiple demographic identities.

Figure 1 illustrates the overall research framework adopted in this study. The research begins with dataset acquisition and preprocessing, followed by computational analysis to identify algorithmic bias and evaluate fairness across demographic groups. The empirical findings are then interpreted using Islamic divine values (tabayyun, ta'aruf, 'adl, and Maqashid al-Shariah) to formulate an integrative framework for ethical AI governance.

Variables and Computational Analysis

The dependent variable was the algorithmic toxicity score, representing the proportion of human annotators who classified a comment as toxic. The score ranges from 0 (non-toxic) to 1 (highly toxic). For classification analysis, the continuous score was binarized using a threshold of 0.5, where scores ≥ 0.5 were categorized as toxic and scores < 0.5 as non-toxic.

The primary independent variables were the identity attributes contained in each comment, including religious identities (Muslim, Christian, Hindu, Buddhist, Jewish, and Atheist), racial/ethnic identities (Black, White, Asian, Latino, and Other Race or Ethnicity), gender identities (Male, Female, Transgender, and Other Gender), sexual orientation (Heterosexual, Homosexual/Gay or Lesbian, and Bisexual), and disability identities (Intellectual or



Source: (Research Result, 2026)

Figure 1. Research Framework

Learning Disability, Physical Disability, and Psychiatric or Mental Illness).

Computational analysis was conducted using Python 3.11, employing pandas 2.0 for data manipulation, NumPy 1.24 for numerical computation, SciPy 1.11 for statistical analysis, and scikit-learn 1.3 for fairness evaluation. The complete analysis scripts are provided in the supplementary materials to ensure research reproducibility.

Data Preprocessing

The dataset initially consisted of **174,716** online comments. Duplicate records were removed prior to analysis. Comments without identity annotations (69,215 observations) were retained for overall toxicity analysis but excluded from identity-based comparisons. The toxicity score was transformed into a binary classification using the 0.5 threshold. To reduce class imbalance, the majority non-toxic class was proportionally undersampled until an approximately balanced distribution (50:50) was achieved, resulting in **90,902** observations. Identity membership followed the dataset convention, whereby an identity annotation score of **≥ 0.5** indicated membership in a particular demographic group. Based on this criterion, the primary comparison groups consisted of 1,396 Muslim-related comments and 1,532 Christian-related comments**.**

Co-occurrence Bias Analysis

To examine contextual bias associated with religious identities, a co-occurrence analysis was performed by measuring the frequency of seven commonly stigmatizing terms—*"terrorist," "radical," "extremist," "bomb," "ISIS," "jihad,"* and *"attack"*—within comments referring to Muslim and Christian identities. The co-occurrence rate was calculated as the proportion of comments in each identity group containing each target term.

Algorithmic Fairness Analysis

Algorithmic fairness was assessed using five widely adopted fairness metrics (Franklin et al., 2024):

1. Demographic Parity Difference the absolute difference in approval rates between a protected group and the reference group. Zero means no disparity; values approaching −1 mean the protected group is maximally disadvantaged.
2. Disparate Impact Ratio the protected group's approval rate divided by the reference group's. Below 0.80 is the threshold employment law and AI fairness research typically use to flag adverse impact (Biddle, 2006).
3. Chi-square test of independence (χ^2) run on the 2×2 contingency table of group membership × moderation outcome, to check whether the approval rate gap was statistically significant.
4. Independent samples t-test applied to mean toxicity scores between groups, with effect size reported via Cohen's d (small = 0.2, medium = 0.5, large = 0.8).
5. 95% confidence interval for the difference in approval rates, computed from the standard error of the proportion difference.

RESULTS AND DISCUSSION

The dataset started at 174,716 online comments. After preprocessing and class balancing, 90,902 observations remained with a balanced class distribution. Table 1 is a general overview of the data.

Tabel 1. General Profile of the Jigsaw Unintended Bias Dataset

Dataset Characteristics	Description
Total Observations (after balancing)	90.902 komentar
Number of Variables	45 kolom (1 ID + 1 teks + 43 atribut)
Target Class Distribution	Balanced: 50% toxic (≥0.5) / 50% non-toxic (<0.5)
Data Time Range	September 2015 – Desember 2017
Number of Identity Groups	24 kelompok (agama, ras, gender, orientasi seksual, disabilitas)
Moderation Status: Approved	74.917 komentar (82,4%)
Moderation Status: Rejected	15.985 komentar (17,6%)

Source: (Research Result, 2026)

Platform records show that 82.4% of comments were approved, while 17.6% were rejected for policy violations. To further examine potential disparities across religious identity groups, fairness metrics were calculated by comparing Muslim- and Christian-identity comments.

Table 2 shows that Muslim-identity comments had an approval rate of 66.3%, compared with 87.4% for Christian-identity comments, representing a 21.1 percentage-point difference. The Disparate Impact Ratio was 0.758, while the chi-square test indicated a statistically significant association between religious identity group and moderation outcome ($\chi^2 = 184.98$, $p < 0.001$). These findings indicate a measurable disparity between the two identity groups within the analyzed dataset.

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 22, No. 2 September 2026 | DOI: 10.33480/pilar.v22i2.8531

Table 2. Fairness Metrics for Muslim and Christian Identity Groups

Fairness Metric	Muslim Identity	Christian Identity	Result / Difference
Number of comments (n)	1,396	1,532	–
Approval Rate	66.3%	87.4%	−21.1 percentage points
Demographic Parity Difference (DPD)	–	–	−0.211*
Disparate Impact Ratio (DIR)	–	–	0.758
Chi-square (χ^2)	–	–	184.98
p-value	–	–	< 0.001

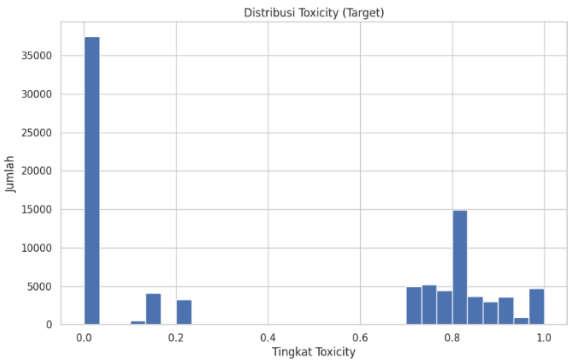
Source: (Research Result, 2026)

Figure 2 presents a word cloud of the most frequently occurring terms in the analyzed comments. The visualization provides an overview of the dominant vocabulary appearing in the Jigsaw dataset and serves as an initial representation of the textual characteristics of the data.



Source: (Research Result, 2026)

Figure 2. Word Cloud of Frequently Occurring Terms in the Jigsaw Dataset

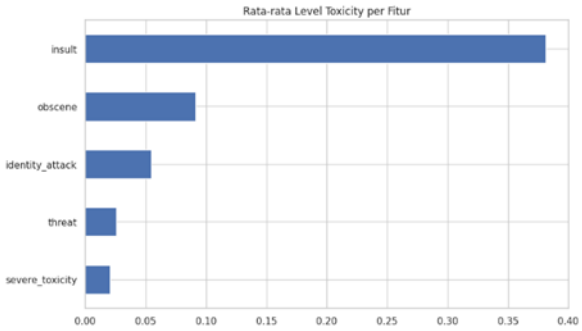


Source: (Research Result, 2026)

Figure 3. Distribution of Toxicity Scores in the Jigsaw Dataset

Figure 3 illustrates the distribution of toxicity scores in the analyzed dataset. The distribution is highly concentrated near the non-toxic range, while a smaller proportion of observations appears in the high-toxicity range. The toxicity distribution is badly skewed. Roughly 37,000 samples cluster near 0 (non-toxic), while only 5,000–15,000 land in the high-toxicity range (0.7–1.0), with a smaller cluster around 0.1-0.25. Left alone, a model trained on this distribution will just favor the non-toxic majority. That's why we applied corrective techniques oversampling, undersampling, class weight adjustment to get the model to actually catch toxic content.

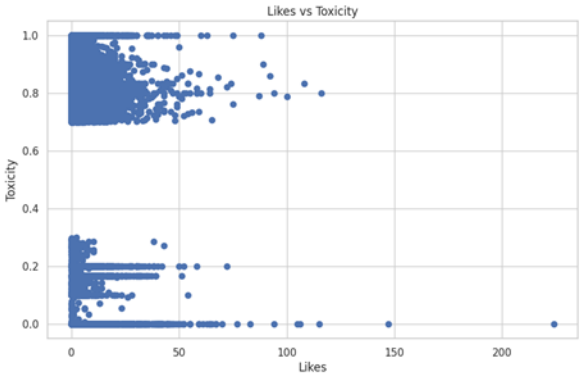
Figure 4 compares the mean scores of the toxicity sub-features in the dataset. Among the analyzed categories, insult shows the highest mean score, followed by obscene and identity attack, whereas threat and severe toxicity occur less frequently.



Source: (Research Result, 2026)

Figure 4. Mean Toxicity Scores Across Toxicity Sub-Features

As shown in Figure 4, insult is the dominant toxicity sub-feature with a mean score of 0.381, substantially higher than obscene (0.092) and identity attack (0.055). In contrast, threat (0.026), sexually explicit content (0.024), and severe toxicity (0.021) show considerably lower mean scores. Insult is clearly the dominant form of toxic speech here which means any model built on this data needs to get good at catching insults specifically, more than the other categories.



Source: (Research Result, 2026)

Figure 5. Relationship Between Likes and Toxicity Scores

Figure 5 shows the relationship between the number of likes received by a comment and its toxicity score. Most observations are concentrated at relatively low numbers of likes across both low and high toxicity scores. As shown in Figure 5, toxicity scores are concentrated primarily within the low (0.0–0.3) and high (0.7–1.0) ranges, while most comments receive fewer than 50 likes. No clear relationship is observed between the number of likes and toxicity scores, indicating that highly toxic and non-toxic comments may receive similar levels of user engagement. Therefore, the number of likes alone does not appear to be a strong indicator of comment toxicity.

Factors Contributing to Algorithmic Bias

Three overlapping problems in the training data explain the bias (Lendvai & Gosztonyi, 2025):

1. Annotation bias: Thousands of volunteers did the annotation, most of them Western and English-speaking. That cultural lens tended to read discussions of Islam and Muslims especially around the post-9/11 and Syrian refugee crisis period (2015–2017) with extra suspicion, which shows up as disproportionate toxicity labeling.
2. Temporal context bias: The dataset comes from 2015–2017, right when anti-Muslim rhetoric on Western social media peaked after a string of terrorist attacks in Europe and the US. That backdrop pushed up how often "Muslim" appeared in negative contexts, and the algorithm picked that up as a toxicity pattern.
3. Co-occurrence bias: Muslim-identity comments appear more frequently alongside negatively connotated terms such as “terrorist,” “radical,” “extremist,” “jihad,” and “attack” than Christian-identity comments. This pattern was directly examined through the co-occurrence analysis, as presented in Table 3.

Table 3. Co-occurrence Rates of Stigmatizing Terms by Religious Identity

Term	Muslim-Identity Comments	Christian-Identity Comments	Difference (percentage points)
Terrorist	13.3%	2.5%	10.8
Radical	5.1%	1.7%	3.4
Extremist	3.2%	1.3%	1.9
Jihad	2.1%	0.5%	1.6
Attack	8.3%	2.9%	5.4

Source: (Research Result, 2026)

As presented in Table 3, all reported stigmatizing terms occurred more frequently in Muslim-identity comments than in Christian-identity comments. The largest difference was observed for the term “terrorist,” which appeared in 13.3% of Muslim-identity comments compared with 2.5% of Christian-identity comments, representing a 10.8 percentage-point difference. Similar patterns were observed for “radical,” “extremist,” “jihad,” and “attack.” These findings indicate systematic differences in the contextual co-occurrence patterns associated with the two religious-identity groups within the analyzed dataset.

Imbalance in Data Representation

Muslim-identity comments in this dataset (n = 1,396, at the ≥ 0.5 group annotation threshold) come in slightly below Christian-identity comments (n = 1,532) despite Islam being the world's second-largest religion. That gap isn't really the issue, though. The dataset was never built to mirror global demographics, so comparing it to world population figures doesn't make methodological sense. What matters more is what's inside each subgroup: were Muslim-identity comments pulled disproportionately from adversarial or conflict-heavy contexts during the 2015-2017 collection window? The co-occurrence numbers above say yes 13.3% terrorist co-occurrence in Muslim comments against 2.5% in Christian ones points to exactly that kind of skew. The model may well have learned identity-linked toxicity patterns that have nothing to do with how these groups are actually discussed day to day (Hosseini, Wieczorek, & Gordijn, 2022).

Limitations of Algorithmic Contextual Understanding

Toxicity detection models built on bag-of-words and TF-IDF the dominant approach when this dataset was collected can't really understand context. A comment that references "Muslim" in a defensive, academic, or journalistic frame gets flagged the same way as one that outright attacks Muslims, just because the keywords overlap.

Part of the problem is structural: classical models score text in isolation, with no view of the conversation around it. That's arguably the deepest weakness in hate speech detection today. Strip away the surrounding discussion and you lose a lot of accuracy without knowing what a comment is replying to, it's genuinely hard to tell if it's hateful. ANAACL 2022 study made this concrete. Researchers built a dataset that classified Reddit comments as hate speech, counter-narrative, or neutral, and when annotators could see the parent comment, 38.3% of their judgments flipped. Nearly two in five comments got reclassified once context entered the picture that's not a number any serious moderation system can shrug off.

And it cuts both ways. A neutral comment can suddenly look hateful once you see what it's replying to; a comment that looks hateful in

isolation can turn out to be counter-speech. Neither mistake is acceptable at scale.

Transformer-based experiments found the same thing from another angle: adding context improved hate speech detection on both binary and multi-label tasks, lifting Macro F1 by 4.2 and 5.5 points respectively. The same exact comment can read as neutral, hateful, or counter-speech depending purely on what came before it something isolated-comment systems simply can't see. Bag-of-words models' blindness to context, in other words, isn't a minor technical gap. It's the ceiling on how good these systems can ever get.

Impact of Digital Discrimination on the Muslim Community

The 21.1-point approval gap ($\chi^2 = 184.98$, $p < 0.001$; Disparate Impact Ratio = 0.758) is statistically significant, and it isn't just a number on a page. If this pattern holds across real platforms, it carries real consequences for how Muslim users experience digital public life. A few implications stand out, though they deserve further empirical testing (Pasipamire & Muroyiwa, 2024):

1. Self-censorship: Muslims who know that talking about their religious identity tends to get flagged will think twice before doing it. The system doesn't need to silence anyone outright the pattern alone does that work.
2. Stigma reinforcement: When an algorithm consistently marks Islamic discourse as more toxic than comparable Christian discourse, it quietly bakes in the assumption that Islam-adjacent content is more dangerous. That assumption then gets replayed across billions of moderation decisions.
3. Unequal participation: Muslim content creators, journalists, academics, and everyday users face a structurally higher risk of moderation than people of other religious backgrounds. The platform isn't neutral ground.

Divine Value Response to Algorithmic Bias

These findings aren't just a technical problem. They touch something more fundamental human dignity, which in the Islamic intellectual tradition falls under *Maqashid al-Shariah*, the core objectives of Islamic law. Imam al-Ghazali and al-Shatibi laid out five such objectives as the highest standard for justice and human welfare. The empirical findings were further interpreted through the five protection dimensions of Maqashid al-Shariah. As presented in Table 4, each dimension provides a normative perspective for examining the potential implications of algorithmic bias on religious expression, human dignity, access to information, cultural continuity, and equal participation (Budiman et al., 2025):

Table 4. Relevance of Maqashid al-Shariah Principles to Algorithmic Bias Findings

Maqashid	Protection Dimension	Relevance to Bias Findings	Form of Violation
Hifzh al-Din (Protection of Religion)	Freedom to practice and express religious identity	Algorithmic bias that disproportionately flags Islamic discourse as toxic creates a climate that suppresses the expression of Muslim religious identity in the digital space	Digital silencing of Islamic expression
Hifzh al-Nafs (Protection of Life)	Human dignity and personal integrity	Algorithmic stigmatization linking Muslim identity with toxicity diminishes the human dignity of the Muslim community	Dehumanization through discriminatory algorithmic labeling
Hifzh al-Aql (Protection of Intellect)	The right to fair and accurate information	Biased algorithms hinder public access to balanced discourse on Islam, as content defending Muslims is more frequently moderated	Information distortion and restriction of intellectual discourse
Hifzh al-Nasl (Protection of Progeny)	Community continuity and cultural value transmission	Systematic silencing of Islamic expression threatens the ability of the Muslim community to transmit values and identity to the next generation in the digital era	Threat to the continuity of cultural-religious identity in the digital space
Hifzh al-Mal (Protection of Wealth/Interests)	The right to equal opportunity and participation	Muslim content creators, journalists, and academics face algorithmic disadvantages that affect the reach of their content and	Digital economic discrimination based on religious identity

Maqashid	Protection Dimension	Relevance to Bias Findings	Form of Violation
		their digital economic opportunities	

Source: (Research Result, 2026)

The Principle of 'Adl (Justice) in Islamic Perspective

'Adl, justice, sits at the foundation of Islamic teaching shaping human relations, social systems, and, by extension, how we govern modern technology. The Qur'an is direct about this. Allah declares:

"إِنَّ اللَّهَ يَأْمُرُ بِالْعَدْلِ وَالْإِحْسَانِ"

"Indeed, Allah commands you to act with justice and to do what is good." (QS. al-Nahl: 90)

Islamic thought doesn't treat justice as satisfied just because identical cases get identical treatment. Substantive *'adl* asks more: attention to real social conditions, lived impact, the specific vulnerabilities each group carries. An AI system can claim formal neutrality all it wants if it produces discriminatory outcomes for a religious community, it fails this standard, intent notwithstanding. This is where algorithmic bias becomes an *'adl* problem directly (Sekar & Ginanjar, 2025). AI models are built from historical data, and that data carries real social, cultural, and religious prejudice. Run it through a system with no corrective mechanism and you don't neutralize the prejudice you scale it up. Every choice along the way training, dataset selection, parameters carries the values of whoever built the system. Neutrality isn't something you get by default. Somebody has to build it in on purpose.

Substantive justice takes active effort, not passive consistency. A moderation system claiming neutrality while producing a 21-point approval gap hasn't met *'adl* it's laundering pre-existing injustice through a technical process. Justice in Islam goes beyond formal equality (treat equals equally); it's substantive, attentive to context and to each party's actual circumstances. For algorithmic bias specifically, 'adl means moderation tools that claim neutrality have to actively find and fix the inequalities they create. Islam rejects identity-based discrimination, full stop which lines up with the *maqāṣid* goals of protecting human dignity *(ḥifẓ al-karāmah)* and promoting the common good *(maṣlaḥah).* Content moderation AI needs both distributive and procedural justice built in, not just technical accuracy.

A 21-point approval gap is about as far from *'adl* as you can get. Calling it "data objectivity" doesn't fix that it just hides it..

The Concept of *Musawah* (Equality) and Its Implications

Islam rejects discrimination based on race, ethnicity, or religion, categorically. The Prophet Muhammad (peace be upon him) said it plainly in the Farewell Sermon: "There is no superiority of an Arab over a non-Arab, nor of a non-Arab over an Arab, nor of a white-skinned person over a dark-skinned person, except by piety."

That's *musawah* equality of human dignity, the idea that all human beings stand equal before God and shouldn't be treated. When AI produces bias against a religious group, it's not just a technical failure. It contradicts justice and equality as Islam defines them, and it runs against universal human rights principles too. At its core, *musawah* is a direct critique of any AI system that ends up applying different standards to different groups, whatever its intent. A global technology platform used by billions of people across every religious and cultural background, operating with systemic bias against one religious group, violates *musawah* on two fronts at once Islamically, and under Articles 2 and 7 of the Universal Declaration of Human Rights.

The Responsibility of Amanah in Technology Development

Amanah, in Islamic thought, isn't reducible to personal honesty. It's a collective responsibility how authority, information, and technology affecting people's lives get managed. Whoever is entrusted with authority, person or institution, owes fair, transparent, accountable conduct. That makes it directly relevant to AI today, since these systems now shape information flow, free expression, and global public opinion at a scale nothing else does. Algorithmic transparency, user rights, bias mitigation these aren't add-ons. They're what Islamic ethics actually demands of AI development.

The weight behind this concept is real. Allah SWT says:

"إِنَّ اللَّهَ يَأْمُرُكُمْ أَنْ تُؤَدُّوا الْأَمَانَاتِ إِلَىٰ أَهْلِهَا"

"God commands you to hand over what is entrusted to you to those it rightfully belongs to." (QS. al-Nisa': 58)

That maps onto technology ethics directly. Developers and companies building AI content moderation systems hold an *amanah* over communication infrastructure that billions of people depend on. Discriminatory outcomes even unintended ones still break that trust.

Our findings, showing differential treatment of comments about Muslims, suggest moderation systems haven't fully lived up to *amanah.* Lower

approval rates for one group over another point to a potential systemic injustice worth taking seriously. And in Islam, *amanah* was never just about developers meaning well it's about actually committing to prevent harm *(ḍarar)* and discrimination.

That framework comes with specific obligations for developers:

1. audit AI systems for bias before deployment, not after the damage is done
2. build training data with real representation across culture, religion, geography, and language
3. bring affected communities especially those marginalized by earlier systems into the design and evaluation process itself
4. keep moderation logic transparent enough for outside audit.

Amanah holds developers accountable for social impact, not just technical performance. In AI moderation that means transparency, institutional accountability, and real effort to stop digital discrimination so the technology actually does the good it claims to.

Transcending Data "Objectivity"

The big myth of the Big Data era is that data is objective, that algorithms run neutral, untouched by politics or social values. They don't. Collection, labeling, selection, interpretation every step runs through people and the social structures they're embedded in. Train an AI system on that data and the biases inside it don't disappear. They get reproduced, often amplified, at automatic, machine speed.

This study's 21-point approval gap makes that concrete. The algorithm isn't anti-Muslim because of some neutral statistical quirk. It's anti-Muslim because the data it learned from was collected during a period soaked in Islamophobic sentiment. The technology absorbed that. Now it plays it back. Comments about Muslims get approved less often than comments about other religious groups which is the algorithm doing exactly what its training data taught it to do, not some glitch in an otherwise neutral system. The bias isn't a property of the technology itself; it's a social construction baked into the dataset, particularly from English-language media and internet discourse during a stretch of time defined by Islamophobic sentiment (Habib, 2025).

Kate Crawford makes a version of this same point in Atlas of AI (2021): AI reflects the society that built it, power relations, inequality, and discriminatory practices included. It's not value-neutral. It's shaped by human politics, economics, culture, ideology all of it. The claim that algorithms hand down objective decisions gets harder to take seriously the closer you look.

Birhane and Prabhu (2021) found something similar: internet-sourced AI datasets routinely misrepresent minority groups, with certain communities overrepresented in negative contexts. Models pick up on that and start associating those identities with risk or threat. Digital discrimination, stacked right on top of the offline kind.

Content moderation specifically gets shaped by which cultures dominate global tech. The companies building these systems are mostly Western, mostly working in English. So "safe" and "problematic" end up defined by a particular cultural lens, onethat doesn't have much room for global religious and cultural diversity. Recent research backs this up automated moderation tends to misread religious discourse, satire, and identity terms used by specific communities, which drives false-positive rates up for minority groups.

Even the idea of "objective" categories falls apart under scrutiny. Deciding what counts as "toxic" or "hate speech" or "unsafe" is a political act, whether anyone admits it or not. Annotators bring their own cultural, religious, and ideological backgrounds to that decision. Gordon et al. (2022) found real variation in how annotators from different backgrounds judged religious and political speech so algorithmic "objectivity" is really just human subjectivity, one layer removed.

Islamic ethics has something useful to say here. The tradition doesn't pretend people are free of prejudice or self-interest quite the opposite which is exactly why it insists every human-made system needs moral oversight. AI doesn't get apass. It can't be an unchallengeable authority. It has to answer to justice *('adl), amanah*, and social responsibility, like everything else people build.

Getting past the myth of data "objectivity" means admitting AI is a socio-political product, not a neutral tool. Anti-Muslim bias in content moderation isn't a bug to patch it's Islamophobia showing up in digital form. Fixing that takes more than better accuracy scores. It means looking hard at where the data comes from, how it gets labeled, who's represented and who isn't, and what the downstream consequences actually are (Razzaq & Razzaq, 2026).

Recommendations for AI System Developers

Four things need to happen:

1. Subgroup-based bias testing should be mandatory in the content moderation development cycle, with experts from potentially affected communities involved from the start.
2. Annotator pools need real diversity across culture, religion, and geography including

Muslim annotators based outside the English-speaking West.

3. Debiasing techniques (resampling, re-weighting, adversarial debiasing) should specifically target the co-occurrence bias this study documented.
4. Explainable AI (XAI) methods should let moderation decisions be publicly audited, not buried inside opaque model weights.

CONCLUSION

Algorithmic bias in content moderation isn't something better models or more data will quietly fix. The real problem runs deeper: AI systems inherit social injustice and then amplify it. "Neutral algorithm" is a claim that doesn't hold up training data always carries the fingerprints of whatever social, political, and cultural moment produced it.

The core finding here is hard to argue with. Comments mentioning Muslim identity got approved 21.1 percentage points less often than comments about Christians 66.3% versus 87.4% ($\chi^2 = 184.98$, $p < 0.001$; Disparate Impact Ratio = 0.758). Not noise. The chi-square confirms it's highly significant, and the Disparate Impact Ratio sits below the 0.8 line used to flag adverse impact. Three reinforcing layers of bias drive that gap annotation bias, historical context bias, co-occurrence bias all plausibly linked to the Islamophobia that soaked English-language internet discourse from 2015 to 2017. The co-occurrence numbers ("terrorist" at 13.3% in Muslim-identity comments versus 2.5% in Christian-identity comments) point to an algorithm that absorbed that pattern and kept reproducing it. One dataset, one time period, though any causal claim needs longitudinal and multi-dataset replication to hold up.

Divine values give this a normative foundation that pure technical analysis can't provide on its own *tabayyun, 'adl, musawah,* the *Maqashid al-Shariah* framework. Through that lens, a system that systematically silences or discriminates against a religious group isn't just behaving unethically. It's contradicting the Qur'anic command to act justly, directly. Amanah pushes this further still: developers are morally responsible for what their systems do to people, not just for how accurate the models are, three things should happen from here. AI developers need to make subgroup-based bias testing mandatory, not optional. National regulators Indonesia especially, with the world's largest Muslim population should require periodic bias audits of platforms operating in their territory. And researchers should build Indonesian-language content moderation datasets that aren't built on Western assumptions, through real collaboration across computer science, Islamic studies, sociology, and law (Marzuki, 2025).

None of this is an argument against AI. It's an argument for being honest about what AI does, and for actually building it better. Fair technology was never going to be perfect technology it's technology that keeps correcting its own mistakes. In Islamic terms, that's what amal saleh, righteous action, looks like in the digital age: using what we know and what we can build for everyone's benefit, not just the majority's.

REFERENCE

Abdul Aziz, M. S., & Zulkepli, M. I. S. (2025). Algorithmic bias in AI-based credit assessment in Islamic financial institutions: A Quranic ethical approach. *QURANICA – International Journal of Quranic Research, 17*(2), 125–156. https://doi.org/10.22452/quranica.vol17no2.6.

Binti Abu Bakar, S., Binti Hussin, H., Bin Ahmad, Z. A. S., & Bin Rozaini, M. A. S. (2025). Reconstructing Artificial Intelligence Ethics through Qur'anic Values of Ihsan and Amanah Using Thematic Exegesis. *International Journal of Quranic Research, 17*(2), 429–446. https://doi.org/10.22452/quranica.vol17no2.18.

Budiman, M., Wijaya, M. M., Rizkillah, R. W., Noor, I. H., Safuan, S., & Destriana, R. (2025). Artificial intelligence (AI) in Islam: Building ethics and solutions based on Tawhid. *Proceeding of the International Conference on Religious Education and Cross-Cultural Understanding, 1*(1), 60–76. https://doi.org/10.61132/icreccu.v1i1.24.

Ferrara, E. (2024). Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. Sci, 6(1), 3. https://doi.org/10.3390/sci6010003

Franklin, G., Stephens, R., Piracha, M., Tiosano, S., Lehouillier, F., Koppel, R., & Elkin, P. L. (2024). The sociodemographic biases in machine learning algorithms: A biomedical informatics perspective. *Life, 14*(6), 652. https://doi.org/10.3390/life14060652.

Habib, Z. (2025). Ethics of artificial intelligence in Maqāṣid al-Sharīʿa's perspective. *KARSA Journal of Social and Islamic Culture, 33*(1), 105–134. https://doi.org/10.19105/karsa.v33i1.19617

He, Y. (2024). Artificial intelligence and socioeconomic forces: Transforming the landscape of religion. *Humanities and Social Sciences Communications, 11*, 602. https://doi.org/10.1057/s41599-024-03137-8.

Hosseini, M., Wieczorek, M., & Gordijn, B. (2022). Ethical issues in social science research employing big data. *Science and Engineering Ethics, 28*, Article 29. https://doi.org/10.1007/s11948-022-00380-7.

Lendvai, G. F., & Gosztonyi, G. (2025). Algorithmic bias as a core legal dilemma in the age of artificial intelligence: Conceptual basis and the current state of regulation. *Laws, 14*(3), 41. https://doi.org/10.3390/laws14030041.

Marzuki, M. F. (2025). Risiko bias algoritma dalam credit scoring berbasis AI pada bank syariah. *Maliki Interdisciplinary Journal, 3*(6), 338–347. https://urj.uin-malang.ac.id/index.php/mij/article/view/15723.

Pasipamire, N., & Muroyiwa, A. (2024). Navigating algorithm bias in AI: Ensuring fairness and trust in Africa. *Frontiers in Research Metrics and Analytics, 9*, 1486600. https://doi.org/10.3389/frma.2024.1486600

Raquib, A., Channa, B., Zubair, T., & Qadir, J. (2022). Islamic virtue-based ethics for artificial intelligence. *Discover Artificial Intelligence, 2*, 11. https://doi.org/10.1007/s44163-022-00028-2.

Razzaq, S., & Razzaq, N. (2026). Artificial intelligence and moral responsibility: An Islamic normative approach. *Journal of Science and Knowledge Horizons, 6*(1), 19–32. https://doi.org/10.34118/jskp.v6i1.4545.

Sekar, D. S., & Ginanjar, A. (2025). Analisis dampak ketergantungan media sosial terhadap motivasi belajar siswa SMP Negeri 1 Bawen. *Tarbiyah bil Qalam: Jurnal Pendidikan Agama dan Sains, 9*(2). https://pdfs.semanticscholar.org/d67a/f9764a36847589d179c5e8fc4ea01e9ded7e.pdf%0Ahttps://pdfs.semanticscholar.org/9a86/3a741413b0468894f6b6f2701092cdc21692.pdf.

Yaakob, M. A., Abdullah, M. R., Mohammad Masruh, M. L., Ku Yaacob, K. M. S., Noor Azman, F. M., Suarni, A., & Hasanudin. (2025). Empowering framework Sunnah values: An ethical framework for research in the age of artificial intelligence (AI). *International Journal of Research and Innovation in Social Science, 9*(29), 44–48. https://doi.org/10.47772/IJRISS.2025.92900008.