DOI 10.18276/978-83-8419-028-9-25

28th European Conference on Artificial Intelligence (ECAI 2025)

THE ETHICAL PARADOX OF AI IN MANAGEMENT: CAN MACHINES MAKE MORAL DECISIONS?

Krystian Redżeb[a] Joanna Sauk[b]

[a] *University of Szczecin, Institute of Management, Szczecin, Poland*
[b] *University of Szczecin, Institute of Spatial Economy and Socio-Economic Geography, Szczecin, Poland*

ABSTRACT

Purpose: *This study examines the ethical challenges of AI-driven decision-making in high-stakes domains such as autonomous vehicles and algorithmic tenant screening. It questions whether AI can act morally or simply optimize for efficiency, assessing its alignment with human ethical values.*

Need for the Study: *As AI systems increasingly shape critical aspects of life—transportation, housing, employment, there is growing concern about their ethical impact. While promising fairness, AI often amplifies societal biases. Existing frameworks lack transparency and safeguards, making governance essential.*

Methodology*: A mixed-methods approach is applied, including case study analysis, large-scale ethical simulations, and cross-cultural comparisons. Empirical data from the MIT Moral Machine experiment and AI-generated moral decisions are analyzed using exploratory factor analysis and hierarchical clustering.*

Findings: *The study reveals sharp differences between human and AI ethics. AI models exhibit stronger biases, favoring younger, wealthier, and law-abiding individuals more than humans. Cultural differences are also evident: Western societies tend toward utilitarianism, while collectivist cultures factor in broader moral dimensions. The SafeRent case showed that AI-driven screening disproportionately disadvantaged minority applicants, reinforcing inequality.*

Practical Implications: *The findings emphasize the need for human oversight in AI systems. Hybrid frameworks that combine algorithmic efficiency with ethical constraints are recommended. Policymakers and business leaders must enforce transparent, culturally aware AI standards that promote fairness. Ethical governance of AI must balance global norms with regional values to ensure equitable outcomes.*

Keywords: Artificial Intelligence; Ethics; Management; Decision-Making; Bias; Governance

Jel Codes: M15, D81, L21

1. INTRODUCTION

The rapid integration of artificial intelligence (AI) into decision-making processes has sparked intense discussions about its ethical implications and practical limitations. AI-driven systems are now employed across various sectors, from healthcare and finance to autonomous transportation and criminal justice, fundamentally reshaping traditional decision-making frameworks. While AI promises efficiency, objectivity, and scalability, its deployment raises pressing ethical concerns, particularly regarding transparency, accountability, and fairness (Etzioni & Etzioni, 2017). The

Peer-review under responsibility of the scientific committee of 28th European Conference on Artificial Intelligence (ECAI 2025) and Intelligent Management Workshop.

ability of AI to make moral decisions remains one of the most contested issues in the field of AI ethics and governance.

One of the most compelling case studies in AI ethics is its application in autonomous vehicles (AVs). These systems must make split-second decisions in potentially life-or-death scenarios, raising questions reminiscent of the classic trolley problem: should an AV prioritize the lives of its passengers over pedestrians? Should it favor younger individuals over the elderly, or law-abiding citizens over jaywalkers? The *Moral Machine* experiment, a large-scale study conducted by MIT, analyzed human ethical preferences in AV decision-making, revealing significant cross-cultural variations in moral judgments (Awad et al., 2018). This study serves as a critical benchmark for understanding the ethical dimensions of AI-driven decision-making and highlights the challenges in developing universally accepted moral frameworks for autonomous systems.

Beyond AVs, AI is increasingly being deployed in areas where ethical considerations are paramount, such as hiring algorithms, financial risk assessments, and legal decision-making. However, empirical evidence suggests that AI systems often amplify existing biases rather than eliminate them. A notable example is the SafeRent algorithm, an AI-driven tenant screening system that disproportionately rejected applicants from marginalized communities due to its reliance on credit scores and non-rental-related financial data (Humber, 2023). This case underscores the risks of algorithmic discrimination and the urgent need for AI governance structures that prioritize fairness and social responsibility.

The growing reliance on AI in high-stakes decision-making necessitates a robust ethical framework that addresses key concerns such as algorithmic bias, decision transparency, and moral alignment. Scholars debate whether AI should replicate human moral intuitions – despite their inherent biases – or adopt an optimized ethical framework based on rationalist principles (Bonnefon, Shariff, & Rahwan, 2016; Jiang et al., 2024). The challenge lies in ensuring that AI-driven decisions align with both societal values and ethical principles while mitigating unintended discriminatory outcomes.

Moreover, the emergence of intelligent management frameworks – where AI systems actively support strategic planning, real-time decision-making, and predictive analytics – raises novel ethical dilemmas. Intelligent management extends beyond automation, integrating AI into complex managerial tasks such as scenario planning, resource allocation, and stakeholder coordination. As these systems evolve, the ethical principles guiding their development must address not only fairness and transparency, but also long-term accountability, strategic foresight, and organizational learning.

This study aims to examine the ethical paradoxes of AI-driven decision-making, focusing on key questions:

- Should AI mirror human moral preferences, or should it be programmed to follow an objective ethical standard?
- How can AI systems avoid reinforcing societal biases in areas such as employment, finance, and autonomous transportation?
- What role should policymakers play in regulating ethical AI development and ensuring accountability?

By addressing these questions, this research contributes to the ongoing debate on AI ethics and governance. Through an analysis of real-world case studies, including the *Moral Machine* experiment and AI-driven tenant screening systems, this study explores how AI can be designed to balance efficiency, fairness, and ethical integrity in decision-making processes. The ultimate objective is to provide insights that inform regulatory frameworks and best practices for the responsible deployment of AI in society.

2. LITERATURE REVIEW

Artificial intelligence (AI) is increasingly involved in decision-making processes with significant ethical implications, particularly in autonomous systems such as self-driving vehicles. Ethical AI development requires an understanding of human moral preferences, which serve as a benchmark for designing AI decision-making frameworks. The Moral Machine experiment (Awad et al., 2018) provides critical insights into how different cultures perceive life-and-death dilemmas in autonomous vehicle (AV) scenarios. This large-scale study gathered human ethical preferences

regarding various factors, including age, social status, law abidance, species, and utilitarian considerations.

2.1 AI and Moral Decision-Making

The integration of AI into ethical decision-making poses unique challenges due to the lack of a universal moral standard. While some scholars advocate for AI systems that closely mirror human ethical choices (Bonnefon, Shariff, & Rahwan, 2016), others argue that AI should adopt an optimized ethical framework, free from human biases (Etzioni & Etzioni, 2017). AI-driven decision-making often emphasizes consistency and efficiency, but it also introduces moral rigidity, as demonstrated in studies comparing AI preferences with human ethical intuitions (Jiang et al., 2024).

2.2 Ethical Frameworks for AI Decision-Making

Various ethical theories have been proposed for AI-driven systems:

- Utilitarianism: AI prioritizes actions that maximize the overall benefit, often favoring the greater number of lives saved (Bentham, 1789).
- Deontology: AI follows predefined moral rules, such as protecting law-abiding citizens over criminals (Kant, 1785).
- Virtue Ethics: AI evaluates decisions based on human-like moral reasoning, prioritizing fairness and equity (Aristotle, 350 BCE).

However, research suggests that AI systems may amplify certain biases, prioritizing younger, wealthier, and law-abiding individuals at significantly higher rates than human decision-makers (Awad et al., 2018). Within intelligent management, these frameworks are applied not just in isolated decision points, but across integrated systems that continuously learn and adapt. For example, AI may dynamically adjust inventory levels, workforce distribution, or risk mitigation strategies based on evolving data—decisions which, while operational, carry ethical consequences. Thus, ethical design must consider both static rules and adaptive behaviors in AI-driven managerial contexts.

2.3 Cross-Cultural Ethical Variations

Human moral preferences are not universal; they differ based on cultural, legal, and societal values. Studies have demonstrated strong regional variations in ethical decision-making (Table 3), with Western cultures favoring utilitarian principles and law abidance, while Eastern cultures demonstrate greater collectivist values (Rahwan et al., 2019). These differences raise a crucial question: Should AI decision-making be standardized globally, or should AI systems be culturally adaptive?

2.4 Bias and Ethical Risk in Intelligent Management

Recent studies have shown that AI models exhibit systematic biases when making moral decisions. A comparative analysis of AI-driven ethical choices and human moral decisions revealed that AI displayed a 90% preference for saving younger individuals, an 83% preference for higher-status individuals, and an 85% preference for law-abiding citizens (Jiang et al., 2024). This raises concerns regarding algorithmic fairness and the extent to which AI should reflect or challenge human biases. In summary, while AI presents opportunities for enhancing decision-making efficiency, it also introduces new ethical dilemmas. The challenge for policymakers and AI developers is to determine how AI should be aligned with human values while avoiding the reinforcement of societal biases.

2.5 AI in Intelligent Management

Intelligent management refers to the integration of AI technologies into strategic and operational decision-making within organizations. It encompasses the use of machine learning, natural language processing, and optimization algorithms to enhance efficiency, responsiveness, and forecasting accuracy. However, embedding AI into management also shifts ethical responsibilities. Decision-makers must ensure that intelligent systems are aligned with corporate values, regulatory requirements, and societal expectations.

3. METHODOLOGY

3.1 Research Design

This study employs a quantitative research approach, analyzing large-scale ethical decision-making data collected from the Moral Machine experiment and AI-driven simulations. The study seeks to:

- Examine human moral preferences in ethical dilemmas.
- Compare AI-driven ethical choices with human decision-making.
- Identify regional and cultural variations in moral decision-making.

3.2 Data Collection

The dataset used in this study consists of two primary sources:

- Moral Machine experiment data (Awad et al., 2018): Over 40 million ethical decisions from 4 million participants across 233 countries.
- AI-generated moral decision simulations (Jiang et al., 2024): Responses from 19 advanced large language models (LLMs) evaluated across six ethical dimensions.

Data was collected using a combination of crowdsourced surveys, simulation-based AI outputs, and hierarchical clustering techniques to identify decision-making patterns.

3.3 Analytical Methods

To evaluate decision-making patterns, the following methods were applied:

- Exploratory Factor Analysis (EFA): Used to determine key ethical decision-making factors.
- Hierarchical Clustering: Applied to group responses into cultural and regional clusters.
- Comparative Analysis: Assessed the differences between AI and human decision-making biases.

Figure 6 provides a visual representation of AI and human ethical preferences, highlighting key disparities in decision-making.

3.4 Ethical Considerations

Given the sensitive nature of ethical decision-making, this study adheres to the following ethical principles:

- Transparency: All data sources are publicly available for verification.
- Fair Representation: Data is analyzed without modifying original decision outcomes.
- AI Bias Awareness: Ethical implications of AI-driven decisions are critically assessed.

3.5 Research Questions

This study explores key managerial and policy-related questions, including:

- Should AI ethics be standardized globally, or should ethical decision-making be adapted to regional values?
- How can AI systems avoid reinforcing human biases in ethical dilemmas?
- What role should policymakers play in defining the moral standards for AI-driven decision-making?

By addressing these questions, this study aims to contribute to the broader discussion on AI ethics and the future of machine-driven decision-making.

4. RESULTS

4.1 Case Study: Ethical Decision-Making in Autonomous Vehicles – Insights from the Moral Machine Experiment

As artificial intelligence increasingly permeates everyday life, autonomous vehicles (AVs) represent one of the most ethically challenging applications. AVs must make split-second decisions in life-or-

death scenarios, raising profound moral and legal questions. The classical *trolley problem* serves as a relevant analogy: should an AV prioritize its passengers over pedestrians, the young over the elderly, or law-abiding individuals over jaywalkers? (Awad et al., 2018). To investigate societal perspectives on such dilemmas, the Massachusetts Institute of Technology (MIT) conducted the *Moral Machine* experiment, a large-scale global study designed to gather human moral preferences on AV decision-making. This study provides critical insights into ethical considerations that should inform AV programming and policymaking (Awad et al., 2018; Bonnefon et al., 2016). The *Moral Machine* experiment utilized an online platform to simulate ethical dilemmas involving unavoidable traffic collisions. Participants were required to decide which groups to save in scenarios that varied across several dimensions, including:

- number of lives at risk (e.g., one pedestrian vs. five passengers),
- age (children vs. elderly individuals),
- social status (homeless individuals vs. professionals),
- law abidance (jaywalkers vs. those crossing at designated areas),
- species (humans vs. animals) (Awad et al., 2018).

This experiment leveraged crowdsourced participation, collecting over 40 million decisions from 4 million respondents across 233 countries and territories. Researchers applied exploratory factor analysis and hierarchical clustering to identify patterns in decision-making and cultural variations (Rahwan et al., 2019). Figure 1 illustrates the percentage of respondents who prioritized saving a human over an animal in life-and-death scenarios. The data highlight how different regions exhibit varying degrees of preference for human life over non-human life.

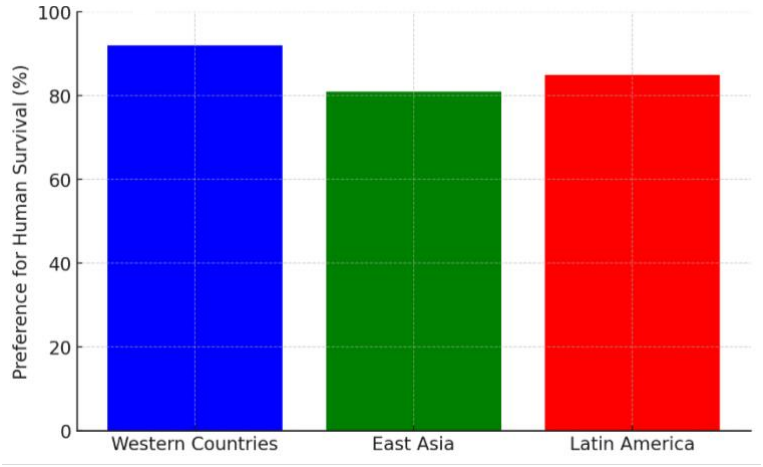

Figure 1. Human vs. Animal Survival Preferences
Source: based on Awad et al. (2018).

In this dataset, 85.6% of respondents globally chose to save a human rather than an animal. Western countries exhibited the strongest preference (~92%), followed by Latin America (85%). East Asian respondents displayed a slightly lower preference (~81%), suggesting a marginally higher ethical valuation of animal life in comparison to other regions **(**Awad et al., 2018).

Figure 2 represents the proportion of respondents who preferred saving the greater number of individuals, regardless of their characteristics. It provides insight into the extent to which people apply a utilitarian approach when making ethical decisions in life-or-death situations.

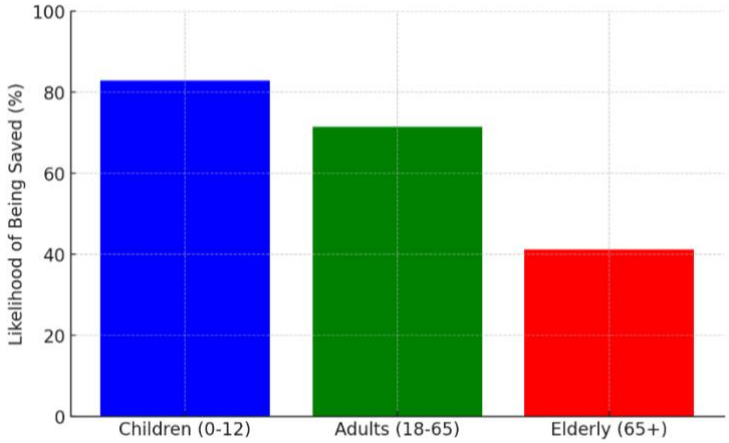

Figure 2. Utilitarian Preference (More Lives Saved)
Source: based on Awad et al. (2018).

Globally, 76.3% of respondents prioritized saving the larger group of individuals. The preference was highest in Western countries (82%), while East Asia and Latin America exhibited slightly lower rates (71–74%). This regional variation may suggest that collectivist cultures incorporate additional ethical considerations beyond sheer numbers when making moral choices (Awad et al., 2018; Rahwan et al., 2019). Figure 3 visualizes the likelihood of different age groups being saved in moral dilemmas. The data show a strong preference for younger individuals, emphasizing the societal value placed on youth over older populations.

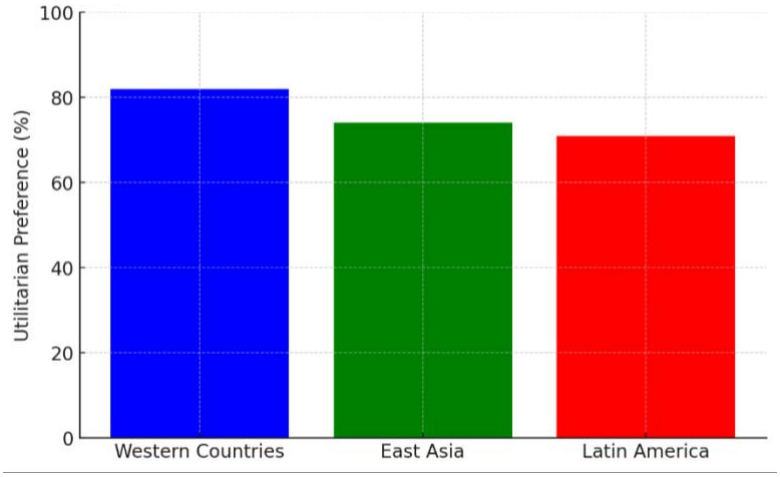

Figure 3. Age-Based Survival Preferences
Source: based on Bonnefon et al. (2016).

The results indicate that children (0–12 years) were saved in 82.9% of cases**,** significantly more often than adults (18–65 years), who had a 71.5% survival rate**.** In contrast, the elderly (65+) were saved in only 41.2% of cases, highlighting a substantial age-based bias in survival decisions (Bonnefon et al., 2016; Jiang et al., 2024). Figure 4 presents the percentage of respondents who prioritized saving law-abiding individuals over those who violated traffic laws. The data illustrate how societal norms regarding rule adherence influence ethical decision-making.

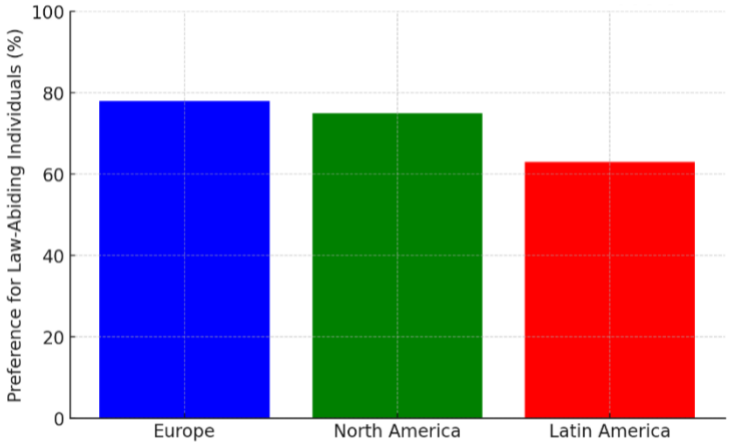

Figure 4. Preference for Law-Abiding Individuals
Source: based on Awad et al. (2018).

Overall, 71.2% of respondents favored saving pedestrians who followed traffic laws**.** This preference was strongest in Europe (78%) and North America (75%), while Latin America exhibited a lower preference (63%). These findings suggest that moral decision-making may be shaped by regional attitudes toward legal and social compliance, which could have implications for AI-driven ethical programming, such as in autonomous vehicles (Awad et al., 2018; Rahwan et al., 2019). This study has critical implications for artificial intelligence (AI) development, especially in systems that must make moral decisions, such as autonomous vehicles, AI-driven healthcare, and crisis management algorithms. Modern AI models often learn from human choices, raising fundamental ethical questions:

- Should AI replicate human morality, given that human decision-making exhibits strong biases based on age, social status, and adherence to laws?
- What ethical norms should be encoded in AI systems when moral decisions vary across cultures and regions?
- Is it fair and ethical for AI to adopt human decision patterns if those patterns favor specific social groups?
- How should managers approach the ethical design of AI systems to minimize discrimination and ensure fair decision-making?

This study underscores the fact that **A**I decisions cannot be made in isolation from societal norms and cultural values. However, AI should also not blindly replicate human biases. Business leaders and policymakers must make deliberate decisions about ethical AI programming, considering both empirical data and social responsibility. A crucial question for managers is:

- What values should guide artificial intelligence to ensure both fairness and societal acceptance?
- Should businesses and policymakers establish universal ethical AI standards, or should AI ethics be tailored to regional moral differences?

This study opens a broader discussion on how AI should make decisions and whether it should mirror human morality or strive to improve it. For decision-makers and technology leaders, this remains one of the most pressing challenges in the ethical implementation of AI. Table 1 presents cross-cultural variations in ethical decision-making preferences observed in the Moral Machine experiment.

The data illustrate how moral choices differ by region, with Western countries exhibiting a strong utilitarian approach, prioritizing the greater number of lives saved and emphasizing law abidance and professional status. In contrast, Eastern cultures demonstrate a more collectivist approach, with a more balanced preference between law-abiding and non-law-abiding individuals and less emphasis on age-based biases. Meanwhile, Latin American and Francophone countries show a stronger preference for higher social standing and display more sympathetic choices toward marginalized individuals and criminals. These cultural differences highlight the complexity of integrating moral decision-making frameworks into AI systems, particularly in global applications like autonomous vehicles (Awad et al., 2018; Rahwan et al., 2019). The stark regional differences in decision-making raise a crucial challenge for AV programming:

- A universal ethical framework would ensure consistency but may not align with cultural values in certain regions.
- A region-specific approach would improve local acceptance but could introduce regulatory and technical complexities.

Table 1. Cross-Cultural Variations in Ethical Preferences.

Cluster	Key Ethical Decision-Making Trends
Western Cluster (North America & Europe)	– Strong preference for saving younger individuals – High emphasis on law abidance and professional status – Most utilitarian decision-making (saving more lives over individual traits)
Eastern Cluster (East Asia & Japan)	– Greater focus on collectivist values, with less emphasis on individual age – More balanced decision-making between law-abiding individuals and others – Slightly lower preference for saving the greater number of individuals than Western counterparts
Southern Cluster (Latin America & Francophone Countries)	– Higher gender biases in decision-making (women were saved 12% more often than men) – Strong preference for saving individuals of perceived higher social standing – More sympathetic choices toward criminals and marginalized individuals than in other regions

Source: based on Awad et al. (2018).

The *Moral Machine* experiment provides invaluable insights into human ethical preferences in autonomous vehicle decision-making. The data confirms several universal trends, such as a preference for saving humans over animals and prioritizing younger individuals. However, significant regional and cultural variations raise complex questions about how AVs should be programmed to reflect ethical diversity. For AVs to gain widespread social acceptance, developers, policymakers, and ethicists must engage in interdisciplinary collaboration, ensuring that AI-driven ethical frameworks align with both societal values and legal standards. Ultimately, the success of AVs depends not only on technical advancements but also on their ability to navigate the profound moral and ethical expectations of human societies.

4.2 In-Depth Case Studies: Ethical and Managerial Challenges in AI-Driven Decision-Making AI Bias in Tenant Screening: The SafeRent Controversy

SafeRent Solutions, formerly known as CoreLogic Rental Property Solutions, utilizes an algorithm-based scoring system to assess prospective tenants' lease performance risk. This system assigns applicants a numerical score ranging from 200 to 800, which landlords use to make rental decisions. The algorithm considers several factors are showed in Table 2.

In 2024, Mary Louis, a security guard from Massachusetts, encountered a significant obstacle in securing housing due to an AI-driven tenant screening system known as SafeRent. Despite her commendable rental history, the algorithm assigned her a low score of 324, leading to an automatic rejection of her application without any avenue for appeal (Humber N.J., 2023).

Table 2. Factors considered in SafeRent Solutions

Factor	Description
Payment History	Timeliness and consistency of past payments
Credit Utilization	Ratio of credit used to credit available
Credit History	Length and status of credit accounts
Credit Availability	Amount of available credit
Inquiries	Number of recent credit inquiries

Source: based on Humber (2023).

Mary Louis had been a reliable tenant for 17 years, consistently paying her rent on time. She possessed a low-income housing voucher, ensuring that a portion of her rent would be covered by government assistance. However, SafeRent's algorithm heavily weighted factors such as credit scores and non-rental-related debts, which did not accurately reflect an applicant's ability to pay rent punctually. This methodology disproportionately affected individuals like Louis, who, despite having some credit card debt, had a proven track record of meeting rental obligations (Khan & Tazel, 2024). This case underscores the necessity for transparency, fairness, and human oversight in AI-driven systems, especially those impacting essential aspects of life such as housing. It highlights the potential for AI algorithms to perpetuate existing biases if not carefully designed and monitored, emphasizing the importance of ethical considerations in AI development and deployment. Monica Douglas, a senior tenant with 26 years of positive rental history, was also subjected to an unfair rejection due to SafeRent's algorithmic decision-making. Although she had previously experienced an eviction, her long-standing record of timely rent payments demonstrated financial responsibility. However, SafeRent assigned her a score below 500, effectively disqualifying her from securing housing. The system failed to account for her decades of rental reliability, reinforcing concerns that AI-driven decision-making models may not be equipped to recognize real-world financial behaviors beyond standardized credit metrics (Humber N.J., 2023).Figure 1 illustrates the disparity in rental application approval rates across racial groups. White applicants experience the highest approval rates, with **81.7%** of applications successfully processed. In contrast, Hispanic applicants receive approvals at a significantly lower rate of **68.5%,** while Black applicants face the highest level of discrimination, securing approvals in only **54.9%** of cases. This substantial gap highlights a systemic issue, as SafeRent's algorithm disproportionately filters out minority applicants. Given that Black and Hispanic populations have historically faced economic marginalization, the reliance on automated systems exacerbates pre-existing inequalities in housing accessibility (Zeng, 2023).

The tenant screening process heavily relies on credit scores, which SafeRent uses as a core determinant of rental eligibility. Figure 2 presents the median credit scores across racial groups, demonstrating significant disparities:

- White applicants have the highest median credit score at 725.
- Hispanic applicants maintain a median score of 661, reflecting a moderate disadvantage in rental application assessments.
- Black applicants possess the lowest median score of 612, placing them at an increased risk of rental rejections (Khan & Tazel, 2024).

This data suggests that the AI-driven screening system inherently disadvantages groups with historically lower access to credit-building financial services, further limiting their housing opportunities.

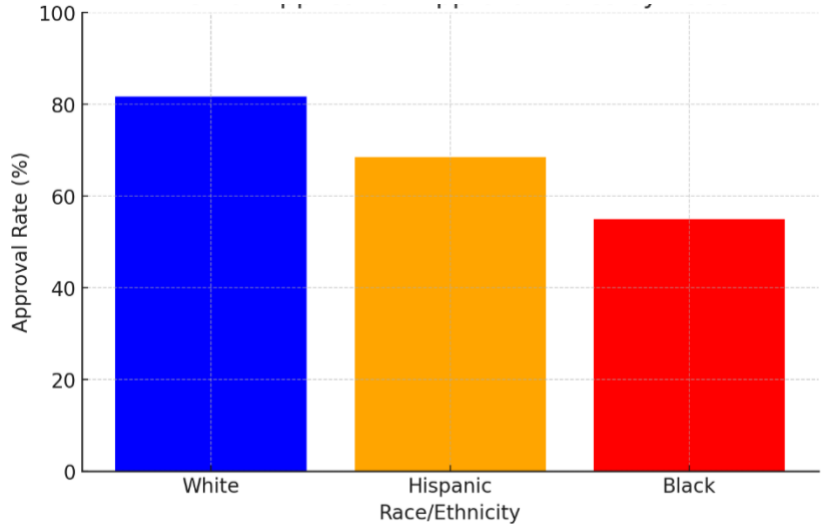

Figure 5. Rental Application Approval Rates by Race
Source: based on Humber (2023).

.

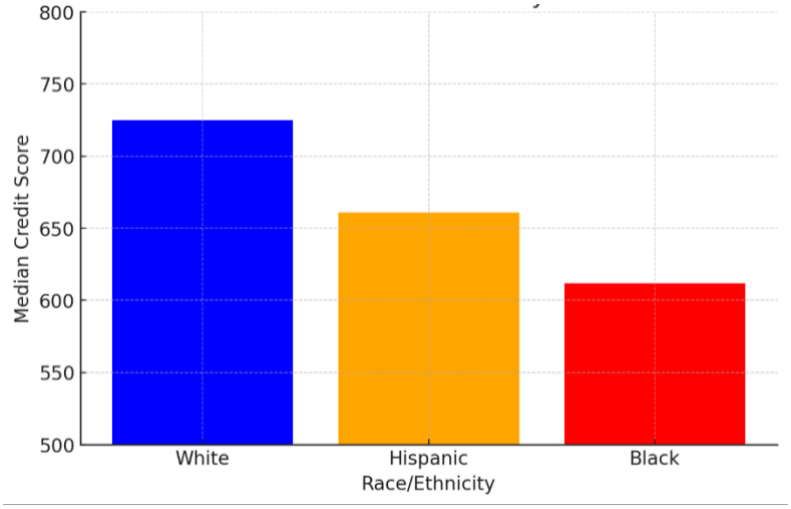

Figure 6. Median Credit Scores by Race
*Source: b*ased on Humber (2023).

Figure 7 displays the percentage of individuals classified as having "subprime" credit scores, defined as scores below 620. This classification significantly impacts rental application outcomes, as SafeRent's algorithm assigns lower rental scores to individuals with subprime credit.

The distribution of subprime credit scores is as follows:

- 45.1% of Black consumers fall into the subprime category, making them the most vulnerable group in AI-driven tenant screening.
- 31.5% of Hispanic consumers also hold subprime credit scores, placing them at a substantial disadvantage in securing rental properties.
- 18.3% of White consumers have subprime credit, significantly lower than the rates observed for Black and Hispanic applicants (Khan & Tazel, 2024).

These statistics highlight how the SafeRent algorithm perpetuates economic disadvantages by overly weighting credit scores, further restricting access to fair housing.

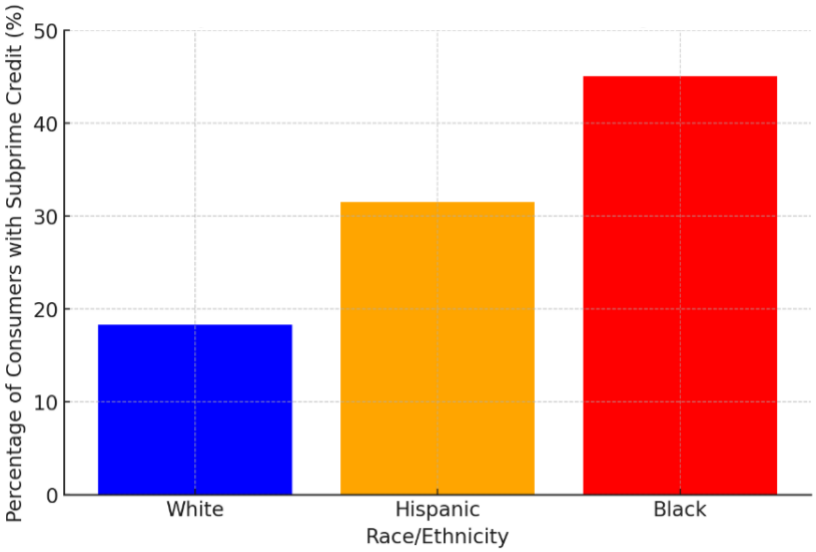

Figure 7. Subprime Credit Score Distribution by Race
Source: based on Humber (2023).

Figure 8 provides a breakdown of the racial demographics of housing voucher holders in Massachusetts, where SafeRent's algorithm was challenged in a 2022 class-action lawsuit. The data reveals that:

- 24% of housing voucher holders are Black,
- 31% are Hispanic,
- 45% belong to other racial and ethnic groups (Humber N.J., 2023).

Given that voucher holders receive partial or full rental assistance from public housing authorities, they should not be subjected to the same financial scrutiny as traditional renters. However, SafeRent's AI system failed to account for the stability provided by government-backed housing vouchers, disproportionately impacting minority groups (Humber N.J., 2023).

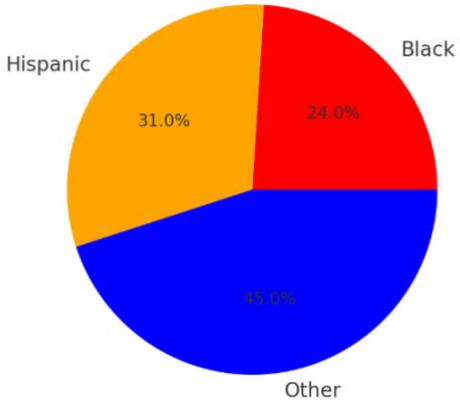

Figure 8. Demographics of Housing Voucher Holders in Massachusetts
Source: based on Humber (2023).

One of the major criticisms of SafeRent's tenant screening system is its failure to recognize the financial stability of voucher holders. Figure 9 illustrates the financial structure of housing voucher payments:

- Public housing authorities cover approximately 73% of rent costs for voucher holders.

- Tenants are responsible for only 27% of the rent, ensuring a stable payment structure for landlords.

By disregarding this factor, SafeRent's scoring model incorrectly categorized voucher holders as high-risk tenants, contributing to widespread rental denials despite their financial backing.

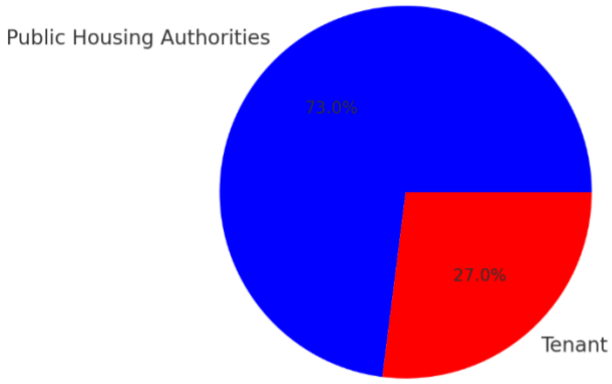

Figure 9. Breakdown of Rent Coverage for Voucher Holders
Source: based on Humber (2023).

In 2022, a class-action lawsuit was filed against SafeRent by over 400 Black and Hispanic tenants in Massachusetts, alleging that the company's algorithm discriminated against them. The plaintiffs argued that the system assigned lower scores to minority renters using housing vouchers compared to white applicants, thereby violating the Fair Housing Act. They contended that the algorithm's reliance on credit information and disregard for the guaranteed payments from housing vouchers led to unjust outcomes (Humber N.J., 2023). The lawsuit culminated in a settlement where SafeRent agreed to pay $2.3 million and cease using its scoring system for applicants with housing vouchers for five years. Additionally, any new scoring models developed by SafeRent must undergo independent validation by a third-party fair housing organization before implementation. This settlement marked a significant step towards addressing biases in AI systems used in critical decision-making processes (Khan & Tazel, 2024).

The SafeRent case underscores the critical need for transparency, fairness, and oversight in AI-driven decision-making systems. It highlights how reliance on certain data points, such as credit history, without considering contextual factors like housing vouchers, can perpetuate systemic biases and result in discriminatory outcomes. As AI continues to permeate various aspects of life, ensuring that these systems promote equity and do not reinforce existing disparities is imperative.

For managers and decision-makers, this case presents critical questions:

- Data Selection and Weighting: How should organizations determine which data points are relevant and how they are weighted in AI algorithms to ensure fair outcomes?
- Transparency and Accountability: What measures can be implemented to make AI decision-making processes more transparent to those affected by them?
- Regulatory Compliance: How can companies ensure that their AI systems comply with existing laws and regulations, such as the Fair Housing Act, to avoid legal repercussions?

Addressing these questions is vital for developing AI systems that are both effective and equitable, thereby fostering trust and reliability in automated decision-making processes.

4.3 Case Study: Moral Preferences of Multilingual Large Language Models in Trolley-like Dilemmas

Recent research conducted by Jiang et al. (2024) explored the ethical alignment of advanced Large Language Models (LLMs) in moral decision-making contexts, specifically through scenarios inspired by the classic trolley problem. The study examined moral preferences of 19 different LLMs, including notable models such as GPT-4, GPT-4o Mini, Llama 3.1, and Gemma 2, across multilingual settings (107 languages). The authors constructed the MULTITP dataset, comprising

precisely 97,520 unique ethical dilemmas, uniformly divided across 107 languages (920 scenarios per language). These vignettes presented moral dilemmas varying systematically along six ethical dimensions: age, gender, fitness, social status, species (human vs. pet), and utilitarian considerations (number of lives saved).

Each model's responses were rigorously evaluated against human baseline data, originally collected from approximately 40 million human decisions from the Moral Machine project (Awad et al., 2018). The researchers employed a quantitative measure—global misalignment scores—defined as aggregated L2-distance between the model and human preferences across all dimensions.

The findings revealed notable numerical differences among models. For instance, the best-aligned model, Llama 3.1 (70B), achieved a relatively low global misalignment score of 0.6, indicating strong similarity to human ethical preferences. Conversely, GPT-4o Mini demonstrated significantly higher misalignment (1.35), suggesting a more rigid or extreme ethical stance. Specifically, GPT-4o Mini displayed substantial preference extremes, presented in Figure 6:

This study was conducted using moral dilemma simulations, where both AI systems and human participants were tasked with making decisions about whom to save in life-threatening scenarios. Human preference data was gathered from large-scale experiments, such as the Moral Machine project, in which participants had to decide who an autonomous vehicle should prioritize in unavoidable accident situations.

Figure 10 presents a comparative analysis of AI-driven and human ethical decision-making preferences in life-and-death scenarios. The data illustrate the extent to which AI models align with or diverge from human moral intuitions across multiple categories, including age, gender, social status, physical fitness, legal compliance, and economic standing.

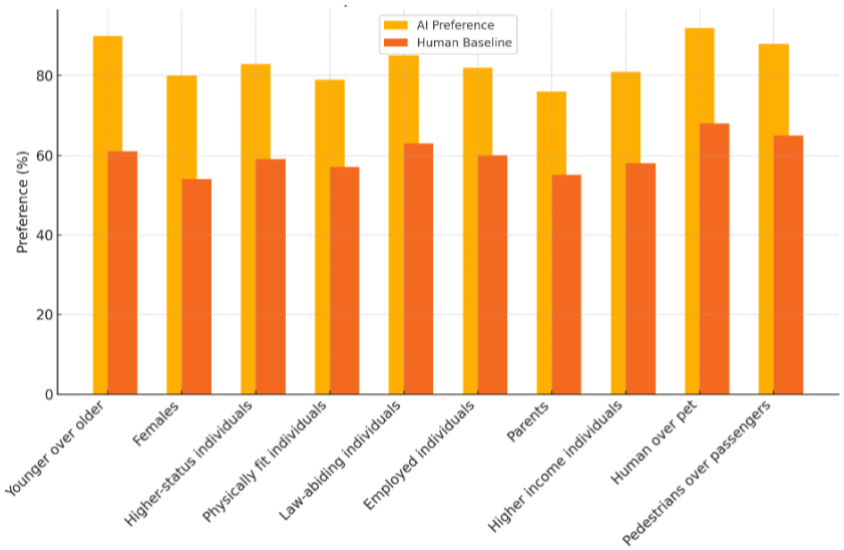

Figure 10. Comparison of AI and Human Preferences
Source: based on Jin et al. (2024).

The analysis compared AI preferences with human decision-making across the following categories:

- Younger individuals over older ones: AI exhibited a 90% preference for saving younger people, whereas humans showed a 61% preference.
- Females over males: AI chose females in 80% of cases, while human preference was 54%.
- Higher-status individuals: AI prioritized them in 83% of cases, compared to 59% for humans.
- Physically fit individuals over those with disabilities: AI displayed a 79% preference, while human preference was 57%.
- Law-abiding citizens over criminals: AI exhibited an 85% preference, whereas humans showed 63%.

- Employed individuals over the unemployed: AI favored the employed in 82% of cases, while humans in 60%.
- Parents over childless individuals: AI saved parents in 76% of cases, compared to 55% for humans.
- Higher-income individuals over lower-income individuals: AI showed an 81% preference, while humans exhibited 58%.
- Humans over pets: AI prioritized humans in 92% of cases, whereas humans did so in 68%.
- Pedestrians over vehicle passengers: AI favored pedestrians in 88% of cases, while humans in 65% (Awad et al., 2018).

This study demonstrates that while AI can make decisions faster and more consistently than humans, its selection criteria can lead to controversial outcomes. The deployment of AI in ethically sensitive areas requires careful calibration to mitigate excessive biases and incorporate ethical reflection that aligns with social values and acceptable moral standards. The analysis highlights that AI's moral decision-making is significantly more rigid and systematic than human choices. This stems from its algorithmic approach, which focuses on specific characteristics rather than applying the flexible and intuitive ethical reasoning inherent to humans.

AI exhibits stronger biases in favor of younger, healthier, and wealthier individuals, raising critical ethical concerns about the implementation of such decision-making models. If AI-driven systems are to be integrated into autonomous vehicles, healthcare decision-making, or crisis management, it will be imperative to incorporate human values and align AI decisions with widely accepted societal norms.

This study demonstrates that while AI can make decisions faster and more consistently than humans, its selection criteria can lead to controversial outcomes. The deployment of AI in ethically sensitive areas requires careful calibration to mitigate excessive biases and incorporate ethical reflection that aligns with social values and acceptable moral standards.

This extensive study raises several critical questions for AI ethics:

- To what extent can – or should – AI developers calibrate multilingual models to culturally or regionally specific moral norms?
- How can AI systems avoid amplifying inherent biases observed in human ethical decision-making, especially given their extreme manifestation in certain LLM outputs?
- What implications do these ethical biases have for practical AI applications, including automated vehicles, healthcare decision support, or global policy-making tools?

5. DISCUSSION

The findings of this study reveal significant ethical challenges in AI-driven decision-making, particularly in the context of autonomous vehicles (AVs). The Moral Machine experiment provides a critical benchmark for understanding how human moral preferences vary across cultures, while AI-driven simulations demonstrate how AI models replicate, amplify, or diverge from these preferences. The concept of intelligent management offers a practical lens through which the ethical paradox of AI becomes operationally relevant. As AI systems assume more responsibility for managerial functions, the need for ethical oversight and explainability grows. Intelligent management is not merely about optimizing KPIs—it is about designing AI ecosystems that act in line with human-centered values, ensuring decisions are not only effective, but also just and inclusive.

One of the most striking observations is that AI systems tend to exhibit stronger biases than human decision-makers. For instance, AI showed a 90% preference for saving younger individuals, whereas human respondents exhibited a significantly lower preference of 61%. Similar trends were observed in AI's preference for saving higher-status individuals (83%) and law-abiding citizens (85%), compared to human preferences of 59% and 63%, respectively. These results raise a critical ethical dilemma: Should AI reflect human moral choices, even when they demonstrate biases, or should AI be programmed to adopt an impartial ethical framework?

Furthermore, the data confirm regional variations in ethical preferences. Western countries demonstrated the highest adherence to utilitarianism (82%), prioritizing the greater number of lives saved. In contrast, East Asia and Latin America showed lower utilitarian tendencies (71–74%),

suggesting that moral decision-making in these regions is shaped by collectivist and social harmony principles. These cultural differences create a major challenge for AI deployment in global applications, as one ethical framework may not align with societal values in every region. AI models trained on human ethical preferences risk inheriting and amplifying societal biases. For example, the SafeRent case study demonstrated how AI-driven tenant screening systems disproportionately affected low-income and minority groups due to their reliance on credit scores and financial indicators. Similarly, AI models used in autonomous vehicles, healthcare, and law enforcement may unintentionally reinforce age, gender, and social status biases, unless explicitly corrected.

This raises a fundamental question for AI ethics:

Should AI be designed to counterbalance human biases, or should it remain a neutral reflection of human morality, even if it results in ethically problematic decisions? To address these concerns, AI governance frameworks must incorporate mechanisms for bias mitigation, transparency, and human oversight. Developing AI systems that balance efficiency, fairness, and ethical accountability will be essential for ensuring public trust and societal acceptance. For business leaders and policymakers, these findings emphasize the urgent need for ethical AI governance. Companies developing AI-driven decision-making systems must consider several key factors:

- Regulatory Compliance: AI systems must align with legal and ethical standards, such as anti-discrimination laws and data protection regulations.
- Transparency & Explainability: AI decisions should be interpretable and justifiable, particularly in high-stakes domains such as healthcare and criminal justice.
- Cross-Cultural Considerations: AI ethics should balance global standardization with regional ethical diversity, ensuring that AI-driven systems are both culturally adaptive and ethically consistent.
- These managerial decisions will play a pivotal role in shaping the future of AI ethics, determining whether AI serves as a neutral arbiter of moral choices or an active agent in promoting fairness and social justice.

6. CONCLUSION

This study underscores the complex ethical challenges posed by AI-driven decision-making, particularly in domains where life-and-death choices must be made. The Moral Machine experiment and AI simulations provide valuable insights into how moral preferences vary across cultures and how AI systems process ethical dilemmas.

The results highlight three major takeaways:

1. AI decisions reflect human biases but often amplify them – AI models exhibited stronger preferences for age, status, and law-abidance compared to human respondents.
2. Cultural differences influence ethical decision-making – Western countries displayed a strong utilitarian tendency, while collectivist cultures considered additional moral dimensions beyond numerical survival rates.
3. AI fairness and transparency remain critical challenges – Without bias mitigation strategies, AI systems risk reinforcing existing inequalities in areas such as healthcare, employment, and transportation safety.

Given these findings, several key challenges and opportunities emerge for AI development:

Should AI ethics be globally standardized or tailored to regional moral preferences?

- How can AI systems be designed to reduce bias without compromising efficiency?
- What role should policymakers play in ensuring AI systems align with societal values and legal frameworks?

Ultimately, the ethical implementation of AI requires collaboration between technologists, ethicists, policymakers, and business leaders. Moving forward, AI governance must strike a balance between technological innovation and ethical responsibility, ensuring that AI serves as a tool for social good rather than a mechanism that reinforces societal inequalities.

For AI to gain widespread public trust and acceptance, its decision-making processes must be transparent, fair, and aligned with widely accepted ethical norms. This study contributes to the ongoing debate on how AI should navigate moral dilemmas, encouraging further research on the intersection of

AI, ethics, and human values. As AI systems increasingly underpin intelligent management practices, ethical governance must evolve in parallel. Addressing these challenges is not only a technical necessity but a managerial responsibility.

Funding: Co-financed by the Minister of Science under the "Regional Excellence Initiative".

REFERENCES

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The Moral Machine experiment. *Nature*, *563*(7729), 59–64. https://doi.org/10.1038/s41586-018-0637-6.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., & Horvitz, E. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*. https://arxiv.org/abs/2303.12712.

Crandall, J. W., Oudah, M., Ishowo-Oloko, F., Abdallah, S., Bonnefon, J.-F., Cebrian, M., Shariff, A., & Rahwan, I. (2018). Cooperating with machines. *Nature Communications*, *9*(1), 233. https://doi.org/10.1038/s41467-017-02597-8

Etzioni, A., & Etzioni, O. (2017). Incorporating ethics into artificial intelligence. *The Journal of Ethics, 21*(4), 403–418. https://doi.org/10.1007/s10892-017-9252-2

Humber, N. J. (2023). A home for digital equity: Algorithmic redlining and property technology. *California Law Review, 111*(October). Retrieved from https://www.californialawreview.org/print/a-home-for-digital-equity

Khan, F., & Tazel, R. (2024). Human rights and AI governance: Addressing the ethical challenges of algorithmic decision-making. Retrieved from: https://www.researchgate.net/publication/383661700_Human_Rights_and _AI_Governance_Addressing_the_Ethical_Challenges_of_Algorithmic_Decision-Making (15.03.2025).

Kirchschläger, P. (2024, September 24). Big Tech firms have consistently shown little concern about harming people and violating their rights. *Le Monde.* Retrieved from: https://www.lemonde.fr/en/opinion/article/2024/09/24/peter-kirchschlager-big-tech-firms-have-consistently-shown-little-concern-about-harming-people-and-violating-their-rights_6727074_23.html (15.03.2025).

Metz, C., & Schmidt, G. (2023, March 29). Elon Musk and others call for pause on A.I., citing 'profound risks to society'. *The New York Times.*

Rahwan, I. (2018). Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology, 20*(1), 5–14. https://doi.org/10.1007/s10676-017-9430-8.

Richards, J. (2023, June 23). Will self-driving cars change moral decision-making? It's time to separate science fact from science fiction about self-driving cars. *Time.* Retrieved from: https://time.com/6179223/self-driving-cars-moral-decision-making/ (15.03.2025).

Shariatmadari, D. (2023, September 2). 'I hope I'm wrong': The co-founder of DeepMind on how AI threatens to reshape life as we know it. *The Guardian*. Retrieved from: https://www.theguardian.com/technology/2023/sep/02/deepmind-co-founder-mustafa-suleyman-ai-reshaping-life (15.03.2025).

Vallor, S. (2023, December 15). Shannon Vallor says AI does present an existential risk—but not the one you think. *Vox.* Retieved from: https://www.vox.com/future-perfect/384517/shannon-vallor-data-ai-philosophy-ethics-technology-edinburgh-future-perfect-50 (15.03.2025).

Zeng, Y. (2023, July 18). UN Security Council meets for first time on AI risks. *Reuters.* Retrieved from: https://www.reuters.com/technology/un-security-council-meets-first-time-ai-risks-2023-07-18/ (15.03.2025).