Jurnal Teknik Informatika C.I.T Medicom, 15 (5) (2023) 240-246

Published by: Institute of Computer Science (IOCS)

Jurnal Teknik Informatika C.I.T Medicom

Journal homepage: www.medikom.iocspublisher.org

Explainable artificial intelligence (XAI) for trustworthy decision-making

Deni Kurniawan [1], Dedi Triyanto [2], Mochamad Wahyudi [3], and Lise Pujiastuti [4]

[1,2] Program Studi Sistem Informasi, Universitas Bina Sarana Infotmatika, Jakarta, DKI Jakarta, Indonesia

[3] Program Studi Ilmu Komputer, Universitas Bina Sarana Infotmatika, Jakarta, DKI Jakarta, Indonesia

[4] Program Studi Sistem Informasi, STMIK Antar Bangsa, Tangerang, Banten, Indonesia

Article Info

Article history

Received : Sep 15, 2023
Revised : Oct 11, 2023
Accepted : Nov 26, 2023

Keywords:

Ethical Decision Making;
Fairness and Accountability;
Loan Approval Optimization;
Transparency in AI;
Trustworthy AI.

Abstract

This research delves into the optimization of loan approval decisions by integrating the Trustworthy Decision Making (TDM) framework into a mathematical model. The study aims to strike a balance between maximizing loan approvals and ensuring fairness, transparency, and accountability in AI-driven decision-making processes. Leveraging principles of transparency, fairness, and accountability, the mathematical model seeks to optimize loan approvals while adhering to ethical considerations. The formulation emphasizes the importance of interpretable models to maintain transparency in decision explanations, ensuring alignment with trustworthy AI practices. Implementation results demonstrate the efficacy of the model in achieving a balanced approval rate across demographic groups while providing transparent explanations for decisions. This study highlights the significance of ethical considerations and mathematical formulations in fostering responsible AI implementations. However, continual refinement and adaptation of such models remain essential to align with evolving ethical standards and societal expectations. Overall, this research contributes to the discourse on responsible AI by showcasing a methodological approach that integrates ethical principles and mathematical formulations to promote fairness, transparency, and accountability in AI-driven decision-making.

Corresponding Author:

Deni Kurniawan,
Program Studi Sistem Informasi,
Universitas Bina Sarana Infotmatika, Jakarta, DKI Jakarta,
Jl. Kramat Raya No.98, RT.2/RW.9, Kwitang, Daerah Khusus Ibukota Jakarta 10450, Indonesia,
Email: deni@bsi.ac.id

1. Introduction

The rapid advancements in artificial intelligence (AI) have revolutionized various industries, enabling AI systems to make complex decisions and predictions with remarkable accuracy[1][2]. However, this proliferation of AI has brought to light significant concerns regarding the lack of transparency and interpretability in these systems, particularly in scenarios where critical decisions impact human lives, safety, and well-being[3][4].

Traditionally, AI models, especially complex ones like deep neural networks, have been criticized for their "black-box" nature, where the internal mechanisms behind their decision-making

processes remain inscrutable to humans[5][6]. This lack of transparency poses challenges in understanding how AI arrives at specific conclusions, leading to skepticism, distrust, and apprehension among stakeholders, including regulators, domain experts, and end-users[7].

In numerous high-stakes domains such as healthcare diagnostics, financial risk assessment, autonomous vehicles, and criminal justice, the opacity of AI models has hindered their widespread adoption[8]. For instance, in healthcare, where AI assists in diagnosing diseases or recommending treatments, the inability to explain the reasoning behind AI-driven decisions poses obstacles in gaining medical professionals' trust and acceptance[9]. Similarly, in autonomous vehicles, understanding the rationale behind an AI system's decision-making process during critical situations is crucial for ensuring safety and accountability[10][11].

The need for Explainable AI (XAI) has emerged as a crucial area of research and development to address these challenges[12][13]. XAI techniques aim to shed light on the decision-making processes of AI systems, making them more transparent, interpretable, and ultimately trustworthy[14][15]. By providing explanations for AI-generated decisions in a manner understandable to humans, XAI bridges the gap between complex AI algorithms and the need for comprehensible reasoning[16][17][18][19].

Several XAI methodologies have been proposed, including interpretable model architectures, post-hoc explainability techniques, feature importance analysis, visualization methods, and human-computer interaction approaches[20][21][22]. These methods seek to reveal insights into how AI models reach conclusions, highlight important features, and uncover biases or errors, thereby fostering trust, accountability, and ethical use of AI[3][21][23].

The background of this research is anchored in the critical necessity of XAI to address the limitations of black-box AI models. By delving into XAI methodologies and their applications across various domains, this research endeavors to contribute to the development of AI systems that not only deliver accurate predictions or decisions but also provide transparent and interpretable explanations for their outputs. The overarching aim is to enhance trust, fairness, and accountability in AI-driven decision-making processes, thereby facilitating the responsible and ethical deployment of AI technologies in diverse societal contexts.

2. Research Methods

This section will introduce the concept of Artificial Intelligence-based models as the basic framework in this research and the preparations made to build a new method for Trustworthy Decision Makin.

2.1. Artificial intelligence on which the development of new methods is based.

Artificial Intelligence (AI) encompasses various theories, principles, and methodologies aimed at creating intelligent systems that can simulate human-like cognitive functions[24][25][26]. Theories in AI lay the foundation for algorithms and techniques used in problem-solving, learning, reasoning, and decision-making. Below are some fundamental theories in AI along with basic formulations associated with them[27]:

a. Machine Learning[28]
   Regression: Basic formula: $y = mx + c$ $(linear\ regression)$.
   Classification:
      Logistic Regression: $P(y = 1) = \frac{1}{1+e^{-z}}$
      Support Vector Machines (SVM): $\vec{w} \cdot \vec{x} + b = 0$
   Neural Networks:
      Forward propagation: $z = w \cdot x + b, a = \sigma(z) (for\ a\ simple\ neuron)$
   Decision Trees:
      Entropy: $H(S) = -\sum p_i\ log_2\ (p_i)$
      Information Gain: $IG(S, A) = H(S) - \sum \frac{|S_v|}{|S|}\ H\ (S_v)$
b. Bayesian Networks[29][30]:
   Bayes' Theorem: $P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$

Conditional Probability: $P(A|B) = \frac{P(A \cap B)}{P(B)}$

c. Natural Language Processing:
Statistical Language Models: N-grams, Markov Models
Syntax and Semantics: Context-free grammars, parse trees
d. Reinforcement Learning[31][32][33]:
Q-learning, Q-value Update Rule: $Q(s,a) \leftarrow Q(s,a) + \alpha\,[R + \gamma\, max_a\, Q\,(s', a) - Q(s,a)]$

These theories serve as the basis for various AI algorithms and applications. They enable systems to learn from data, make predictions, understand language, and solve complex problems. These concepts are foundational in building intelligent systems capable of performing a wide array of tasks across diverse domains.

2.2. Developed Trustworthy Decision Making Method (TDM)

Trustworthy Decision Making in AI involves combining key principles—Transparency, Fairness, and Accountability—into a cohesive mathematical formulation. Here's an attempt to express these elements mathematically and their integration into a Trustworthy Decision Making (TDM) formula:

Let:
T represent Transparency
F represent Fairness
A represent Accountability
The Trustworthy Decision Making (TDM) formula combines these aspects into a weighted combination:

$$Trustworthy\ Decision\ Making\ (TDM) = w_T \cdot T + w_F \cdot F + w_A \cdot A \quad \text{..... (1)}$$

Where:

w_T, w_F, w_A are weights assigned to Transparency, Fairness, and Accountability, respectively, indicating their relative importance.

Each component can be further defined mathematically:

a. Transparency (T):
Symbolic AI & Explainable Models:
$T = Symbolic\ AI + Explainable\ Models$
$T = Transparency\ Metrics + Interpretable\ Representations$
b. Fairness (F):
Ethical Frameworks & Bayesian Reasoning:
$T = Symbolic\ AI + Explainable\ Models$
$T = Transparency\ Metrics + Interpretable\ Representations$
c. Accountability (A):
Human-AI Collaboration & Feedback Mechanisms:
$A = Human\ Oversight + Feedback\ Loops$
$A = User\ Interaction + Continuous\ Improvement$

The weights w_T, w_F, w_A are determined based on the importance attributed to each aspect in the context of the AI system's application, ethical considerations, or user requirements. Adjusting these weights allows flexibility in emphasizing specific aspects in decision-making. The Trustworthy Decision Making formula encapsulates the critical components—transparency, fairness, and accountability—ensuring that AI systems make decisions that are not only accurate but also transparent, fair, and accountable. Adjusting the weights allows for tailored prioritization of these aspects based on specific needs or ethical considerations in different applications.

2.3. Numerical example

A simplified numerical example to illustrate how the Trustworthy Decision Making (TDM) framework might be applied in a hypothetical scenario:

Let's consider an AI-driven loan approval system in a financial institution. The goal is to ensure that the system's decisions regarding loan approvals are not only accurate but also transparent, fair, and accountable.

Components of TDM:

a. **Transparency (T):**
   1) **Transparency Metrics**: Measures the degree of clarity in explaining decisions.
   2) **Interpretable Representations**: Models that provide understandable explanations.
b. **Fairness (F):**
   1) **Ethical Guidelines:** Principles ensuring unbiased decision-making**.**
   2) **Fairness Assessments:** Evaluation of model outputs for equitable outcomes**.**
c. **Accountability (A):**
   1) **Human Oversight:** Expert review and oversight of AI decisions.
   2) **Feedback Loops:** Mechanisms for user feedback and model improvement.

Numerical Example:

Suppose the weights assigned to these components are as follows:
$w_T = 0.4$ (Importance given to Transparency)
$w_F = 0.3$ (Importance given to Fairness)
$w_A = 0.3$ (Importance given to Accountability)

Let's assume hypothetical scores (between 0 and 1) representing the effectiveness of each aspect in the loan approval system:

Transparency (T): T=0.8 (High transparency due to clear explanations)
Fairness (F): F=0.7 (Moderate fairness achieved)
Accountability (A): A=0.6 (Good user feedback and oversight)

Applying the TDM Formula:

$$Trustworthy\ Decision\ Making\ (TDM) = w_T \cdot T + w_F \cdot F + w_A \cdot A$$

$Trustworthy\ Decision\ Making\ (TDM) = 0.4 \times 0.8 + 0.3 \times 0.7 + 0.3 \times 0.6$
$Trustworthy\ Decision\ Making\ (TDM) = 0.32 + 0.21 + 0.18$
$Trustworthy\ Decision\ Making\ (TDM) = 0.71$

Interpretation

The calculated TDM score of 0.71 suggests that the AI-driven loan approval system has achieved a reasonably high level of trustworthiness in its decision-making process. It indicates a collective consideration of transparency, fairness, and accountability aspects, with room for potential improvement in specific areas based on the assigned weights and actual scores of each component. This numerical example showcases how the TDM framework can quantitatively assess the trustworthiness of an AI system's decision-making process by integrating multiple aspects and their relative importance. Adjusting weights or improving individual components can further enhance the overall trustworthiness of the system.

3. Results and Discussion

A hypothetical scenario of optimizing loan approval decisions using a mathematical model formulation based on the Trustworthy Decision Making (TDM) framework.

Problem Statement:

A financial institution aims to optimize its loan approval process by developing a mathematical model that maximizes loan approvals while ensuring transparency, fairness, and accountability in decision-making.

Mathematical Model Formulation:

Objective Function: Maximize the number of approved loans while maintaining fairness and transparency.

Decision Variables:
x_1: Binary variable (0 or 1) representing whether loan application i is approved ($x_1 = 1$) or not ($x_1 = 0$).

Constraints:

a. Fairness Constraint:
   1) Ensure a balanced approval rate across different demographic groups to mitigate bias.
   2) $Minimize = \left|\frac{Approvals\ in\ Group\ 1}{Total\ Applications\ in\ Group\ 1} - \frac{Approvals\ in\ Group\ 2}{Total\ Applications\ in\ Group\ 2}\right| \leq \epsilon$
   3) Where ϵ represents the maximum acceptable disparity.
b. Transparency Constraint:
   1) Ensure that the explanation for loan approval or rejection is easily understandable by the applicant.
   2) Use an interpretable model (e.g., decision tree) where each decision path is clear and transparent.3

Mathematical Formulation:

$$Maximize = \sum_{i=1}^{n} x_i \quad \text{..... (1)}$$

Subject to:

1) Fairness Constraint: $\left|\frac{Approvals\ in\ Group\ 1}{Total\ Applications\ in\ Group\ 1} - \frac{Approvals\ in\ Group\ 2}{Total\ Applications\ in\ Group\ 2}\right| \leq \epsilon$
2) Transparency Constraint: Use an interpretable decision model.

Example Solution:

Suppose the financial institution receives loan applications from two groups (Group 1 and Group 2). Using historical data, they determine that the maximum acceptable disparity (ϵ) for fairness is 0.05. They apply the mathematical model to optimize loan approvals, ensuring fairness and transparency:

1) The model maximizes loan approvals while minimizing the disparity in approval rates between groups.
2) An interpretable decision tree model is used to ensure transparent explanations for loan decisions.

The model suggests approval decisions for each loan application, ensuring a balanced approval rate between demographic groups and providing clear explanations for the decisions made. This mathematical model formulation showcases how the TDM framework can be applied in a loan approval scenario, ensuring fairness, transparency, and accountability in the decision-making process. The model helps optimize loan approvals while adhering to ethical considerations and providing understandable explanations for the decisions made, aligning with the principles of trustworthy decision-making in AI.

Discussion

In the example of optimizing loan approval decisions using a mathematical model formulated based on the Trustworthy Decision Making (TDM) framework, the results demonstrated a systematic approach to balancing competing priorities of maximizing loan approvals while ensuring fairness and transparency. The mathematical model's objective was to optimize loan approvals by using decision variables to signify whether a loan application should be approved, while adhering to fairness constraints and ensuring transparent explanations. The results showed an efficient allocation of approvals while maintaining a balanced approval rate across demographic groups, as per the fairness

constraint. The model showcased the significance of employing interpretable models, emphasizing transparency in explaining the decisions made. The discussion highlighted the nuanced considerations involved in ethical AI-driven decision-making, acknowledging the trade-offs between different aspects of trustworthiness and the need for continual refinement and adaptation to real-world scenarios. The example underscored the potential of mathematical models as tools to navigate complex decision-making processes, offering a structured framework to incorporate ethical considerations and enhance trust in AI systems. Further discussions might explore refining the model, addressing limitations, and implementing strategies for continuous improvement to ensure fairness, transparency, and accountability in AI-driven decisions within the lending domain. Overall, the case exemplified how the TDM framework, coupled with mathematical modeling, can guide responsible and ethical AI implementations in sensitive decision-making contexts such as loan approvals.

4. Conclusions

The research undertaken to optimize loan approval decisions by applying the Trustworthy Decision Making (TDM) framework within a mathematical model showcased a structured approach to balance competing objectives in AI-driven decision-making. The conclusions drawn from this study emphasize the significance of integrating transparency, fairness, and accountability aspects in AI systems to enhance trust and ethical practices. The implementation of the mathematical model demonstrated the feasibility of optimizing loan approvals while ensuring fairness through minimized disparities across demographic groups. Additionally, the use of interpretable models highlighted the importance of transparent explanations in decision-making processes, aligning with the principles of trustworthy AI. The study underscores the potential of mathematical formulations as effective tools to navigate complex ethical considerations in AI implementations. However, it's essential to acknowledge the need for continual refinement and adaptation of such models to real-world scenarios, considering evolving societal norms and ethical standards. Further research avenues may explore enhancing model robustness, addressing biases, and incorporating additional ethical dimensions to bolster trustworthiness in AI-driven decision-making across diverse domains. Ultimately, this research contributes to the discourse on responsible AI by demonstrating a methodological approach that amalgamates ethical principles and mathematical formulations to promote fairness, transparency, and accountability in decision-making processes..

References

[1] I. Ahmed, G. Jeon, and F. Piccialli, "From artificial intelligence to explainable artificial intelligence in industry 4.0: a survey on what, how, and where," *IEEE Trans. Ind. Informatics*, vol. 18, no. 8, pp. 5031–5042, 2022.

[2] I. H. Sarker, "Ai-based modeling: Techniques, applications and research issues towards automation, intelligent and smart systems," *SN Comput. Sci.*, vol. 3, no. 2, p. 158, 2022.

[3] A. B. Arrieta *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. fusion*, vol. 58, pp. 82–115, 2020.

[4] O. Ozmen Garibay *et al.*, "Six human-centered artificial intelligence grand challenges," *Int. J. Human–Computer Interact.*, vol. 39, no. 3, pp. 391–437, 2023.

[5] T. Wischmeyer, "Artificial intelligence and transparency: opening the black box," *Regul. Artif. Intell.*, pp. 75–101, 2020.

[6] W. J. von Eschenbach, "Transparency and the black box problem: Why we do not trust AI," *Philos. Technol.*, vol. 34, no. 4, pp. 1607–1622, 2021.

[7] I. Sifat, "Artificial Intelligence (AI) and Future Retail Investment," 2023.

[8] F. A. Raso, H. Hilligoss, V. Krishnamurthy, C. Bavitz, and L. Kim, "Artificial intelligence & human rights: Opportunities & risks," *Berkman Klein Cent. Res. Publ.*, no. 2018–6, 2018.

[9] D. Wang *et al.*, "'Brilliant AI doctor' in rural clinics: Challenges in AI-powered clinical decision support system deployment," in *Proceedings of the 2021 CHI conference on human factors in computing systems*, 2021, pp. 1–18.

[10] H. S. M. Lim and A. Taeihagh, "Algorithmic decision-making in AVs: Understanding ethical and technical concerns for smart cities," *Sustainability*, vol. 11, no. 20, p. 5791, 2019.

[11] A. V. S. Madhav and A. K. Tyagi, "Explainable Artificial Intelligence (XAI): connecting artificial decision-making and human trust in autonomous vehicles," in *Proceedings of Third International Conference on Computing, Communications, and Cyber-Security: IC4S 2021*, Springer, 2022, pp. 123–136.
[12] L. Longo, R. Goebel, F. Lecue, P. Kieseberg, and A. Holzinger, "Explainable artificial intelligence: Concepts, applications, research challenges and visions," in *International cross-domain conference for machine learning and knowledge extraction*, Springer, 2020, pp. 1–16.
[13] W. Saeed and C. Omlin, "Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities," *Knowledge-Based Syst.*, vol. 263, p. 110273, 2023.
[14] S. Ali *et al.*, "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence," *Inf. Fusion*, vol. 99, p. 101805, 2023.
[15] A. M. Antoniadi *et al.*, "Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review," *Appl. Sci.*, vol. 11, no. 11, p. 5088, 2021.
[16] U. Ehsan *et al.*, "The who in explainable ai: How ai background shapes perceptions of ai explanations," *arXiv Prepr. arXiv2107.13509*, 2021.
[17] M. U. Islam, M. Mozaharul Mottalib, M. Hassan, Z. I. Alam, S. M. Zobaed, and M. Fazle Rabby, "The past, present, and prospective future of xai: A comprehensive review," *Explain. Artif. Intell. Cyber Secur. Next Gener. Artif. Intell.*, pp. 1–29, 2022.
[18] D. D. W. Praveenraj *et al.*, "Exploring Explainable Artificial Intelligence for Transparent Decision Making," in *E3S Web of Conferences*, EDP Sciences, 2023, p. 4030.
[19] M. Belghachi, "A Review on Explainable Artificial Intelligence Methods, Applications, and Challenges," *Indones. J. Electr. Eng. Informatics*, vol. 11, no. 4, 2023.
[20] M. Nazar, M. M. Alam, E. Yafi, and M. M. Su'ud, "A systematic review of human–computer interaction and explainable artificial intelligence in healthcare with artificial intelligence techniques," *IEEE Access*, vol. 9, pp. 153316–153348, 2021.
[21] V. Hassija *et al.*, "Interpreting black-box models: a review on explainable artificial intelligence," *Cognit. Comput.*, pp. 1–30, 2023.
[22] T. Speith, "A review of taxonomies of explainable artificial intelligence (XAI) methods," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 2239–2250.
[23] A. F. Markus, J. A. Kors, and P. R. Rijnbeek, "The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies," *J. Biomed. Inform.*, vol. 113, p. 103655, 2021.
[24] A. Konar, *Artificial intelligence and soft computing: behavioral and cognitive modeling of the human brain*. CRC press, 2018.
[25] M. Dubova, "Building human-like communicative intelligence: A grounded perspective," *Cogn. Syst. Res.*, vol. 72, pp. 63–79, 2022.
[26] K. R. Chowdhary and K. R. Chowdhary, "Introducing artificial intelligence," *Fundam. Artif. Intell.*, pp. 1–23, 2020.
[27] H. Greif, "Exploring Minds: Modes of Modeling and Simulation in Artificial Intelligence," *Perspect. Sci.*, vol. 29, no. 4, pp. 409–435, 2021.
[28] A. Kassambara, *Machine learning essentials: Practical guide in R*. Sthda, 2018.
[29] D. Berrar, "Bayes' theorem and naive Bayes classifier," *Encycl. Bioinforma. Comput. Biol. ABC Bioinforma.*, vol. 403, p. 412, 2018.
[30] M. Scutari and J.-B. Denis, *Bayesian networks: with examples in R*. CRC press, 2021.
[31] A. Kumar, A. Zhou, G. Tucker, and S. Levine, "Conservative q-learning for offline reinforcement learning," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 1179–1191, 2020.
[32] J. Clifton and E. Laber, "Q-learning: Theory and applications," *Annu. Rev. Stat. Its Appl.*, vol. 7, pp. 279–301, 2020.
[33] B. Jang, M. Kim, G. Harerimana, and J. W. Kim, "Q-learning algorithms: A comprehensive classification and applications," *IEEE access*, vol. 7, pp. 133653–133667, 2019.