World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

Exploring the balance between explainability and automation in AI-enabled data policy enforcement systems

Adedayo Hakeem Kukoyi *

Purdue University, Department of Information Technology- Data Analytics, West Lafayette, Indiana, United States of America.

World Journal of Advanced Research and Reviews, 2023, 20(03), 2429–2434

Publication history: Received on 17 October 2023; revised on 20 December 2023; accepted on 28 December 2023

Article DOI: https://doi.org/10.30574/wjarr.2023.20.3.2423

Abstract

Adopting artificial intelligence (AI) technologies for data governance and compliance monitoring has changed how organizations implement their data policies. At the same time, technological innovations create tensions between explainability, understanding the reasoning behind the AI's decisions, and automation, which focuses on streamlining processes to the greatest extent possible. This research aims to address the question of potential trade-offs between the two competing goals of AI-assisted systems for data policy enforcement. A quantitative research design was used to sample 100 respondents from data management, cybersecurity, and AI governance communities, and responses were interpreted using descriptive statistics (frequency and percentage). Findings show that although 70% of respondents see assurance of compliance as a reason for automated systems to be transparent and interpretable, 65% regard automation as the primary goal that most automation frameworks should achieve. Results point to the need to hybrid governance frameworks to XAI (explainable artificial intelligence) approaches that may be needed to weave into governance automated frameworks. The study offers suggestions for balancing ethical explainability and operational blockade dismissal in AI system automation frameworks.

Keywords: Explainable AI; Automation; Data Policy Enforcement; Governance; Transparency; Accountability

1. Introduction

AI has transformed how we handle automation, particularly in monitoring, decision-making, and compliance for data management and enforcement policies. Systems with AI capabilities now routinely detect data policy breaches, monitor access control, and enforce organizational data policies in real time (Leslie, 2021). Nevertheless, with automated integration, the challenge of transparency arises: the more a system operates independently, the more difficult it becomes to explain how it works (Floridi, 2023).

AI governance involves explainability and automation. Explainability allows a system to be transparent, auditable, and consistent with human reasoning. In contrast, automation increases efficiency by reducing human control and allowing the system to operate at machine speed (Jobin, Ienca and Vayena, 2019). Therefore, optimal regulation is necessary for compliance, credibility, and trust in AI governance systems.

This research examines how organizations understand and manage this balance. It examines the implications of prioritizing explainability for the rate and precision of automated policy enforcement, and how automation sits alongside human interpretability. This research addresses the ethical AI governance gap arising from the coexistence of automation and human interpretability. It proposes solutions to balance efficiency and transparency in the enforcement of data policies.

* Corresponding author: Adedayo Hakeem Kukoyi.

2. Literature review

The accelerated implementation of Artificial Intelligence (AI) into data governance structures has improved how organizations administer, protect, and enforce data policies. Nonetheless, the incorporation of AI into governance has created a complex dilemma between explainability and automation that must be resolved. This dilemma must be addressed to prevent AI data policy enforcement systems from becoming ineffective and to ensure they remain unclouded, fair, and compliant with legislation (Tzimas, 2023).

2.1. The Concept of AI Explainability in Data Governance

Trust and understanding the outcomes and predictions made by AI models is called explainability (Doshi-Velez and Kim, 2017). In explainability, AI focuses on decisions regarding access control, compliance, data sharing, and the impacts of automated AI tools on data governance. Miller (2019) argued that, from a human accountability perspective, explainability and the accountability of complex computational reasoning are linked. Without sufficient interpretability, automated data enforcement mechanisms will likely become "black boxes" - systems whose decisions are insurmountable and impossible to audit.

Regarding AI technologies, automated data policy enforcement systems incorporate the organizational principles of fairness, transparency, and accountability. Ribeiro, Singh, and Guestrin (2016) promote LIME and other techniques regarding post-hoc modeling tools. These ideas assist Data Officers and Auditors in understanding the reasoning behind a particular enforcement decision, fostering confidence in compliance with the EU General Data Protection Regulation (GDPR) and Nigeria’s Data Protection Act.

2.2. Automation and Efficiency in Policy Enforcement

Within AI governance, automation is the application of machine learning and autonomous agents to oversee, review, and enforce compliance with data policies with minimal human involvement (Gasser and Almeida, 2017). The scaling of automation technologies and the ability to respond to situations instantly are significant advantages for any organization. The implementation of manual control and data policy enforcement is no longer realistic given the rapidly increasing volumes of data from distributed systems, cloud services, and IoT devices. Time saved is enormous when policy violations are automatically detected by AI, self-corrected, and self-reported, with operational reporting executed within seconds.

On the other hand, the complete absence of a manual control mechanism is dangerous. Binns (2018) points out the dangers of automated enforcement systems and how they may deliver on the expected functional requirements but, from a human policy perspective, may deliver utter nonsense, such as disallowed data access that is essential for legitimate use. Automation must be equipped with an oversight mechanism that provides, and documents, reasonable punitive and non-punitive enforcement options and balances human automation with guided automation.

2.3. Balancing Explainability and Automation

The roots of tension between explainability and automation lie in a trade-off between a system's interpretability and a process's efficiency. Deep learning systems automate processes with predictive accuracy, but complex systems are also the most difficult to interpret. Simpler systems can provide explanations but are too slow and inflexible to enforce policies in real time, as needed in many situations.

Arrieta et al. (2020), and especially Lipton (2018), propose that a workable compromise will require hybrid configurations that combine automated systems with XAI components. These configurations enable dynamic control, allowing the system to make low-risk, repetitive decisions and automate the process, while passing higher-stakes or ambiguous situations to a human. This system of cooperation, human-in-the-loop AI governance, provides the desired balance between efficiency and accountability.

2.4. Ethical and Regulatory Implications

The ethical and regulatory ramifications of the intersection of explainability and automation are also profound. Particularly in the context of enforcing a data policy, one must consider the implications of decisions involving personal data. Possible negative consequences of such decisions include discrimination, privacy violations, and reputational damage. The governance of AI must consider ethics and the demands of transparency, equity, and respect for human dignity. The fundamental essence of automated systems that cannot be explained is the violation of the ethical demand of accountability for the consequences of mistakes.

Regulatory concerns also broaden the implications of automation when explainability is not provided. Global provisions, including the General Data Protection Regulation (GDPR) and the ISO/IEC 38505 family of standards, emphasize algorithmic transparency and the need for human oversight in automated decision-making. Provisions at the national level, such as the Nigeria Data Protection Regulation (NDPR), affirm the right of individuals to know the 'why' and 'how' of their data being processed and to be involved in automated decision-making. The legal and ethical expectations for explainable automation are clear.

2.5. Technological Innovations and Emerging Approaches

The latest advancements in technology aim to minimize the gap between automation and explainability. For instance, in Explainable Reinforcement Learning (XRL), policy enforcement agents justify their actions in real time (Puiutta and Veith, 2020). In addition, causal inference frameworks have begun to be used in automated decision-making to enhance the interpretability of decision outcomes (Pearl and Mackenzie, 2018). Furthermore, novel frameworks such as trustworthy AI and ethical-by-design systems aim to prioritize explainability as a fundamental design principle rather than an afterthought (Dignum, 2019).

This innovation makes way for self-explanatory automation. AI systems will be built to automatically provide explainable, traceable decision-making pathways. This holds the promise of a radical shift in AI-enabled policy enforcement systems, from opaque executors to accountable and collaborative partners in the governance of data.

2.6. Challenges and Research Gaps

Though developed systems have made progress, some issues remain unresolved. The first is the lack of a global standard that defines what 'explainability' means (Vilone and Longo, 2021). The next challenge is the widening gap between the technical and managerial perspectives on AI decisions, which complicates the integration of AI governance. Finally, jurisdictional, cultural, and institutional diversity affects the interpretation of 'acceptable automation', making the alignment of governance standards even more complex.

In the future, automation and explanation-balancing scales will need to be used in context-sensitive frameworks. The overlap between automated reasoning and human cognitive models will be a significant step towards explainable automated systems that can provide traceable, understandable justifications for their actions.

3. Methodology

3.1. Research Design

A quantitative descriptive design was used for this study, along with structured questionnaires to collect primary data from professionals engaged in AI deployment, data security, and compliance.

3.2. Population and Sampling

In this case, purposive sampling was used to select 100 respondents, which included AI engineers, data protection officers, IT auditors, and policy analysts from both private tech companies and regulatory bodies.

3.3. Instrument for Data Collection

There were 20 closed-end items under 3 themes for the questionnaire:

- Awareness and understanding of AI explainability
- Importance of automation in data policy enforcement
- Balance between explainability and automation

3.4. Method of Data Analysis

Responses were analyzed using descriptive statistics, specifically, frequencies and percentage distributions. The results were displayed in tables and followed by an explanation.

4. Results

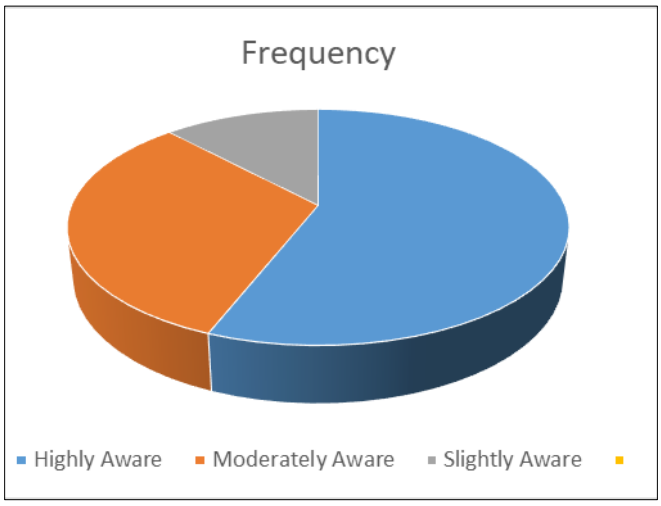

Figure 1 Awareness of Explainable AI (XAI) Concepts

Most respondents (56%) recognize the concepts of XAI, indicating that explainability is an important part of AI ethics. Although 12% are slightly aware, this shows the need for further education and training on AI interpretability.

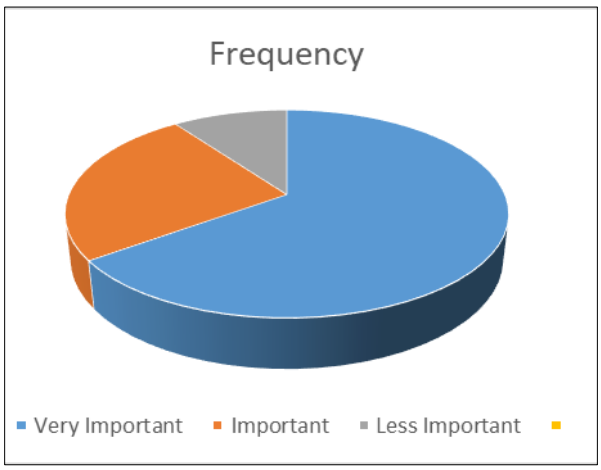

Figure 2 Importance of Automation in Policy Enforcement

Most respondents (65%) view automation as very important for data policy enforcement, reflecting industry expectations for real-time compliance and efficiency. This preference may reduce human oversight of the policy, thereby improving interpretability.

Table 1 Preferred Balance Between Explainability and Automation

Preference	Frequency	Percentage (%)
More Emphasis on Explainability	30	30
Equal Balance	45	45
More Emphasis on Automation	25	25
Total	100	100

Almost half (45%) of the respondents said that there is an equal balance between explainability and automation. This shows that respondents believe both governance frameworks should carry equal weight. 25% preferred automation, while 30% preferred explainability, suggesting compromise is possible.

5. Discussion of Findings

Findings indicate that while automation takes precedence, explainability is more operationally prioritized. Respondents state that enforcement systems become ineffective without automation, but trust and legitimacy are lost without explainability. The inclination towards a more balanced proposition (45%) supports previous literature suggesting hybrid models (Floridi et al., 2022; OECD, 2023).

As Jobin et al. (2019) indicate, organizations are increasingly incorporating interpretability tools into automated systems. Thus, the data supports the view that explainability is not opposed to automation, but rather a design feature that could enable performance while ensuring compliance accountability.

6. Conclusion

This study examines the equilibrium between automation and explainability in AI-enforced systems for data policy. Explainability enhances legal compliance and ethical accountability while automation improves efficiency and scalability. The combination of interpretability and automation as a unified approach is likely to be the most effective for maintaining operational confidence and public trust. Thus, this study advances the integration of explainable AI into automated systems is essential for responsible and transparent data governance.

Recommendations

- Balance automated governance with human supervision to adopt hybrid models.
- Integrate tools for model interpretation with reasoning behind AI actions to promote transparency.
- Expand training in automation technologies for staff, along with ethical AI governance, to augment automated systems.
- Regulators should mandate explainability standards for all automated compliance systems as policy reform.
- Evaluate automation's effects on accountability regularly to promote ongoing compliance with governance structures.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R. and Herrera, F., (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities, and challenges toward responsible AI. Information Fusion, 58, pp.82–115. 10.1016/j.inffus.2019.12.012

[2] Binns, R. (2018). Fairness in Machine Learning: Lessons from Political Philosophy. Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency, pp.149–159. https://proceedings.mlr.press/v81/binns18a.html

[3] Dignum, V., (2019). Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way. Springer Nature, Cham. 10.1007/978-3-030-30371-6

[4] Doshi-Velez, F. and Kim, B. (2017). Towards a Rigorous Science of Interpretable Machine Learning. arXiv preprint arXiv:1702.08608.

[5] Floridi, L., (2023). The Ethics of Artificial Intelligence: Principles, Challenges and Opportunities. AI and Society, 38(1), pp.1–14. https://doi.org/10.1093/oso/9780198883098.001.0001

[6] Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., and Vayena, E. (2018). AI4People—an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), 689–707. https://doi.org/10.1007/ s11023-018-9482-5.

[7] Gasser, U. and Almeida, V. A. F., (2017). A Layered Model for AI Governance. IEEE Internet Computing, 21(6), pp.58–62. https://doi.org/10.1109/MIC.2017.4180835

[8] Jobin, A., Ienca, M. and Vayena, E. The global landscape of AI ethics guidelines. Nat Mach Intell 1, 389–399 (2019). https://doi.org/10.1038/s42256-019-0088-2.

[9] Leslie, D. (2021). Understanding Artificial Intelligence Ethics and Safety: A Guide for the Responsible Design and Implementation of AI Systems in the Public Sector. The Alan Turing Institute. DOI:10.48550/arXiv.1906.05684

[10] Lipton, Z. C., (2018). The Mythos of Model Interpretability. Communications of the ACM, 61(10), pp.36–43. https://doi.org/10.1145/3233231

[11] Miller, T., (2019). Explanation in Artificial Intelligence: Insights from the Social Sciences. Artificial Intelligence, 267, pp.1–38. https://doi.org/10.1016/j.artint.2018.07.007

[12] OECD (2022), "OECD Framework for the Classification of AI systems", OECD Digital Economy Papers, No. 323, OECD Publishing, Paris, https://doi.org/10.1787/cb6d9eca-en.

[13] Pearl, J. and Mackenzie, D., (2018). The Book of Why: The New Science of Cause and Effect. (1st. ed.). Basic Books, Inc., USA.

[14] Puiutta, E. and Veith, E. M., (2020). Explainable Reinforcement Learning: A Survey. In: 2020 IEEE 18th International Conference on Machine Learning and Applications (ICMLA). IEEE, pp.1–9. DOI:10.1007/978-3-030-57321-8_5

[15] Ribeiro, M. T., Singh, S. and Guestrin, C., (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[16] Themistoklis Tzimas (2023), 29, European Public Law, Issue 4, pp. 385-411, https://kluwerlawonline.com/journalarticle/European+Public+Law/29.1/EURO2023021.

[17] Vilone, G. and Longo, L., (2021). Notions of Explainability and Evaluation Approaches for Explainable Artificial Intelligence. Information Fusion, 76, pp.89–106. https://doi.org/10.1016/j.inffus.2021.05.009