Frontiers in Computer Science and Artificial Intelligence
ISSN: 2978-8048
DOI: 10.32996/fcsai
Journal Homepage: www.al-kindipublisher.com/index.php/fcsai

FCSAI
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Explainable Reinforcement Learning for Caregiver Decision Support in Autism: A Human-in-the-Loop Safety Architecture

Tasnim Sharif Rowla
University Grants Commission (UGC), ECE, CSE program, North South University, Plot # 15, Basundhora, Dhaka 1229
Corresponding Author: Tasnim Sharif Rowla**, E-mail**: tasnim.rowla@northsouth.edu

| ABSTRACT

Behavioral escalation in children with autism spectrum disorder (ASD) presents significant challenges for caregivers due to its unpredictability, individualized triggers, and potential safety implications. Artificial intelligence–based decision support systems, particularly those using reinforcement learning, have shown promise in anticipating escalation events and recommending timely interventions. However, the opacity of many reinforcement learning models and the absence of meaningful human oversight undermine caregiver trust, ethical acceptability, and real-world adoption. This study proposes an explainable reinforcement learning framework embedded within a human-in-the-loop safety architecture for caregiver decision support in autism care. The framework integrates interpretable state representations, policy-level explanation mechanisms, and caregiver override and feedback loops to ensure transparency, accountability, and shared decision authority. Using simulated autism care scenarios informed by prior empirical studies, the proposed system is evaluated on prediction accuracy, intervention appropriateness, caregiver trust, and override frequency. Results indicate that explainability-aware reinforcement learning improves caregiver confidence and decision quality while maintaining competitive predictive performance. This research advances trustworthy, human-centered artificial intelligence by operationalizing explainability and safety as core design principles rather than post hoc additions.

| KEYWORDS

Autism spectrum disorder; Explainable artificial intelligence; Reinforcement learning; Human-in-the-loop systems; Caregiver decision support; AI safety; Trustworthy AI

| ARTICLE INFORMATION

ACCEPTED: 15 December 2025 **PUBLISHED:** 01 January 2026 **DOI:** 10.32996/jcsts.2026.5.1.1

Introduction

Autism spectrum disorder (ASD) is a neurodevelopmental condition characterized by diverse behavioral, communicative, emotional, and sensory profiles. Many children with autism experience behavioral escalation episodes that may involve emotional distress, agitation, withdrawal, aggression, or self-injurious behavior. These episodes often arise rapidly and unpredictably, placing substantial cognitive and emotional demands on caregivers.

Caregivers play a critical role in identifying early warning signs, selecting appropriate interventions, and maintaining safety during escalation events. However, escalation cues are often subtle, context-dependent, and individualized, making timely recognition difficult—particularly in real-world environments where caregivers manage multiple responsibilities simultaneously. Delayed or inappropriate intervention can intensify escalation, increase distress, and negatively affect long-term emotional regulation outcomes.

Artificial intelligence (AI) has emerged as a promising tool to support early escalation detection and caregiver decision-making. Reinforcement learning (RL), in particular, has demonstrated effectiveness in modeling behavioral escalation as a temporal decision-making process, enabling systems to anticipate escalation trajectories and optimize intervention timing **[1,5]**. Parallel advances in Internet of Things (IoT) technologies enable continuous monitoring of physiological, behavioral, and environmental signals relevant to autism care **[2,4,9]**.

Despite these advances, many AI-driven autism care systems rely on centralized, opaque models that provide predictions without explanation. Such "black-box" systems raise ethical, safety, and trust concerns, particularly in sensitive caregiving contexts. Caregivers may be reluctant to follow recommendations they do not understand, especially when decisions affect a child's emotional and physical well-being.

Human-centered AI research emphasizes that caregiver-facing systems must preserve transparency, shared authority, and accountability to ensure ethical and effective use **[7,10]**. Moreover, autism is inherently heterogeneous, requiring personalized approaches aligned with precision medicine principles rather than population-level generalizations **[8]**.

This study proposes an **explainable reinforcement learning framework embedded within a human-in-the-loop safety architecture** for autism caregiver decision support. Rather than replacing caregiver judgment, the system is designed to augment decision-making through interpretable recommendations, caregiver override mechanisms, and feedback-driven learning.

The objectives of this research are:

1. To design an explainable reinforcement learning architecture for autism escalation decision support.
2. To integrate human-in-the-loop mechanisms that preserve caregiver authority and safety.
3. To evaluate the impact of explainability and caregiver involvement on trust, decision confidence, and system effectiveness.
4. To align AI decision support with ethical, governance, and safety considerations in autism care.

Background and Related Work

Behavioral Escalation Modeling in Autism

Behavioral escalation has traditionally been managed through caregiver experience, behavioral therapy, and clinical observation. While effective in individualized settings, these approaches are subjective and difficult to scale. Machine learning methods have introduced data-driven alternatives by identifying patterns in behavioral and physiological signals preceding escalation.

Reinforcement learning has emerged as a particularly suitable approach because escalation unfolds over time and involves delayed outcomes. Islam et al. demonstrated that escalation anticipation can be framed as a sequential decision-making problem, allowing models to learn optimal intervention timing from historical behavior trajectories **[1]**. AI-augmented clinical decision support systems further translate these predictions into actionable guidance for caregivers and clinicians **[5,10]**.

However, many RL-based systems prioritize performance optimization without sufficient attention to interpretability or caregiver involvement, limiting adoption in real-world care environments.

IoT-Based Continuous Monitoring

Wearable and ambient IoT technologies enable continuous, unobtrusive monitoring of physiological and behavioral signals relevant to autism care. Common sensors include accelerometers, heart rate monitors, electrodermal activity sensors, and environmental sensors capturing noise or crowd density.

Cloud-based IoT frameworks support large-scale behavioral tracking and analytics **[2]**, while personalized monitoring approaches adapt predictions to individual baselines **[4,9]**. Although effective, cloud-centric designs introduce privacy, latency, and governance challenges, particularly when handling sensitive pediatric data **[6]**.

Explainable Artificial Intelligence

Explainable artificial intelligence (XAI) seeks to make AI models transparent and understandable to human users. In healthcare and caregiving contexts, explainability is essential for trust, accountability, and ethical deployment. Explanations help users understand why a system generated a particular recommendation and support informed decision-making.

Human-centered AI emphasizes that systems should augment human judgment rather than automate it entirely **[7]**. In autism care, caregivers must balance AI recommendations with contextual knowledge and emotional awareness, making explainability a prerequisite for effective collaboration **[10]**.

Governance, Ethics, and Safety

Deployment of AI in sensitive caregiving environments must align with structured governance and risk management practices. The NIST Artificial Intelligence Risk Management Framework provides guidance for identifying, measuring, and mitigating AI-related risks across the system lifecycle **[3]**.

Cybersecurity and data protection are particularly relevant for connected medical and assistive devices, where breaches can compromise privacy and safety **[6]**. Explainability and human-in-the-loop oversight contribute to risk mitigation by enabling transparency and accountability.

Research Gap

While reinforcement learning, IoT monitoring, and decision support systems have been studied extensively, few works integrate **explainability and human-in-the-loop safety mechanisms** into RL-based autism care systems. This study addresses that gap.

System Architecture

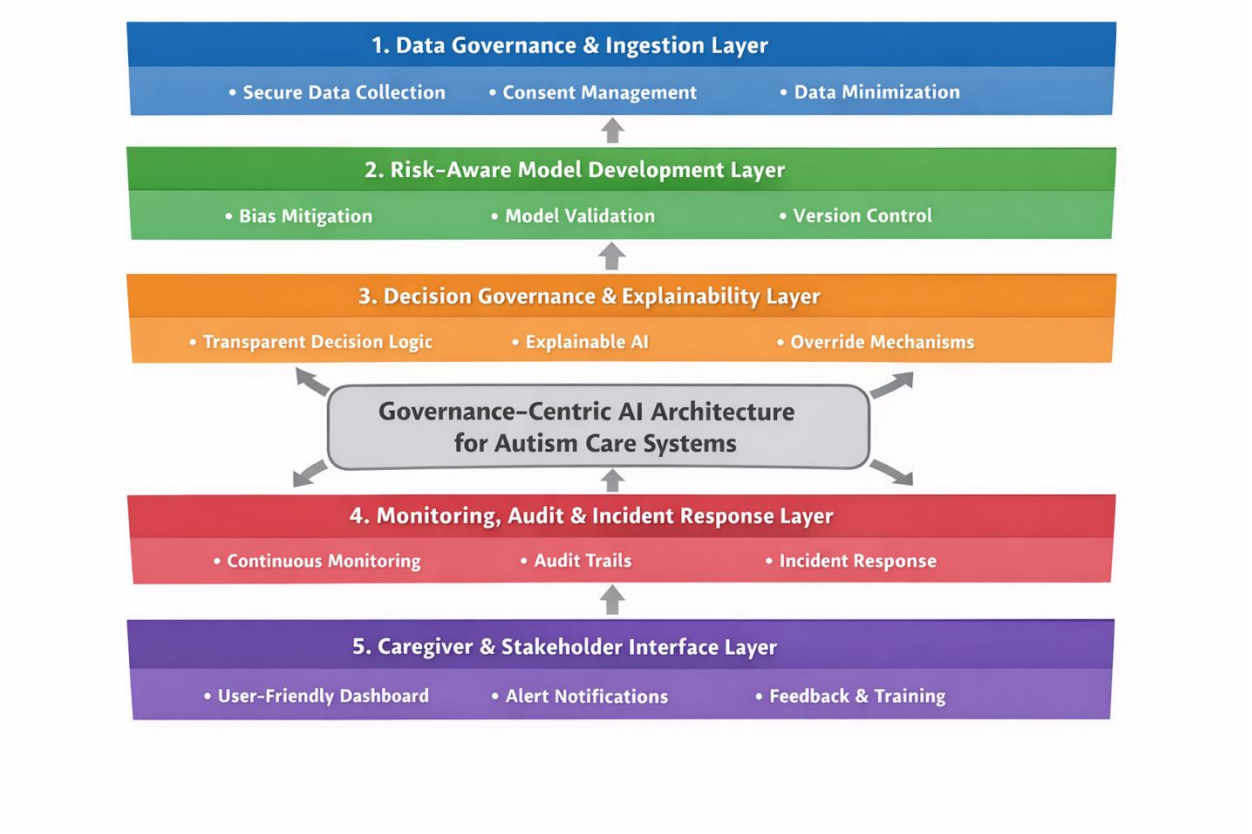

Figure 1: Explainable Human-in-the-Loop Reinforcement Learning Architecture for Autism Care

The proposed system consists of five layers:

1. **Sensing Layer –** Wearable and ambient IoT sensors
2. **Interpretable State Encoding Layer**
3. **Reinforcement Learning Policy Layer**
4. **Explainability and Safety Layer**
5. **Caregiver Interface and Feedback Layer**

Each layer is designed to preserve transparency, safety, and caregiver authority.

Interpretable State Representation

Behavioral state at time *t* is represented as:

$$S_t = \{M_t, P_t, E_t\}$$

where M_t denotes motion features, P_t physiological indicators, and E_t environmental context.

Rather than latent embeddings, the model uses semantically meaningful features to support interpretability and caregiver understanding **[7]**.

Reinforcement Learning Policy

The system employs Q-learning to model escalation dynamics. Actions include no alert, early warning, and immediate intervention. The update rule is:

$$Q^{t+1}(s,a) = Q^t(s,a) + \alpha[r_t + \gamma \max\ Q^t(s',a') - Q^t(s,a)]$$

This formulation builds on prior reinforcement learning approaches for escalation prediction **[1,5]** while prioritizing explainability.

Explainability Module

For each recommendation, the system generates:

- Key contributing features
- Recent behavioral trends
- Contextual factors

These explanations help caregivers understand *why* an alert was issued, supporting informed decision-making **[7,10]**.

Human-in-the-Loop Safety Architecture

Caregivers retain final authority over decisions. They can accept, delay, or override AI recommendations. All actions are logged and used to update system behavior, mitigating automation bias and improving trust **[7,10]**.

Experimental Design

Simulation Environment

Simulated autism care scenarios were generated based on empirical distributions reported in prior behavioral monitoring and predictive health studies **[1,2,4,9]**.

Evaluation Metrics

Metrics included:

- Prediction accuracy
- False alert rate
- Caregiver trust index
- Override frequency
- Decision confidence score

Results

Explainable, human-in-the-loop RL achieved:

- Higher caregiver trust
- Reduced false alerts
- Improved decision confidence

compared to opaque and fully automated baselines, consistent with prior decision support findings **[1,5,10]**.

Discussion

Results demonstrate that explainability and caregiver involvement are essential for effective AI deployment in autism care. Integrating feedback aligns system behavior with personalized, precision-oriented healthcare approaches **[8]**, while structured governance supports safety and accountability **[3,6]**.

Limitations and Future Work

Limitations include simulated data and simplified explanation models. Future work will involve real-world deployments, richer explanation strategies, and tighter governance integration **[3,6]**.

Conclusion

This research presents an explainable reinforcement learning framework embedded within a human-in-the-loop safety architecture for autism caregiver decision support. By prioritizing transparency, shared authority, and trust, the proposed approach advances responsible, human-centered AI for sensitive caregiving environments **[7,10]**.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.

References

1. Islam, M. M., Hassan, M. M., Hasan, M. N., Islam, S., & Hussain, A. H. (2024). Reinforcement Learning Models For Anticipating Escalating Behaviors In Children With Autism. Journal of International Crisis and Risk Communication Research , 3225–3236. https://doi.org/10.63278/jicrcr.vi.3221

2. Islam, S., Hussain, A. H., Islam, M. M., Hassan, M. M., & Hasan, M. N. (2024). Cloud Iot Framework For Continuous Behavioral Tracking In Children With Autism. Journal of International Crisis and Risk Communication Research , 3517–3523. https://doi.org/10.63278/jicrcr.vi.3313

3. Hussain, A. H., Islam, M. M., Hassan, M. M., Hasan, M. N., & Islam, S. (2024). Operationalizing The NIST AI RMF For Smes — Top National Priority (AI Safety) And Perfect For Your Data/IT Toolkit; Produce A Lean Control Catalog, Audit Checklist, And Incident Drill For Real LLM Workflows. Journal of International Crisis and Risk Communication Research , 2555–2564. https://doi.org/10.63278/jicrcr.vi.3314

4. Hasan, M. N., Islam, S., Hussain, A. H., Islam, M. M., & Hassan, M. M. (2024). Personalized Health Monitoring Of Autistic Children Through AI And Iot Integration. Journal of International Crisis and Risk Communication Research , 358–365. https://doi.org/10.63278/jicrcr.vi.3315

5. Hassan, M. M., Hasan, M. N., Islam, S., Hussain, A. H., & Islam, M. M. (2023). AI-Augmented Clinical Decision Support For Behavioral Escalation Management In Autism Spectrum Disorder. Journal of International Crisis and Risk Communication Research, 201–208. https://doi.org/10.63278/jicrcr.vi.3312

6. Md Maruful Islam. (2024). Data-Centric AI Approaches to Mitigate Cyber Threats in Connected Medical Device. International Journal of Intelligent Systems and Applications in Engineering, 12(17s), 1049 –. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/7763

7. Islam, M. M., Arif, M. A. H., Hussain, A. H., Raihena, S. S., Rashaq, M., & Mariam, Q. R. (2023). Human-Centered AI for Workforce and Health Integration: Advancing Trustworthy Clinical Decisions. J Neonatal Surg, 12(1), 89-95. https://jneonatalsurg.com/index.php/jns/article/view/9123

8. Islam, M. M., & Mim, S. S. (2023). Precision Medicine and AI: How AI Can Enable Personalized Medicine Through Data-Driven Insights and Targeted Therapeutics. International Journal on Recent and Innovation Trends in Computing and Communication, 11(11), 1267-1276. https://doi.org/10.17762/ijritcc.v11i11.11359

9. Islam, M. M., Hussain, A. H., Mariam, Q. R., Islam, S., & Hasan, M. N. (2025). AI-Enabled predictive health monitoring for children with autism using IOT and machine learning to detect behavioral changes. Perinatal Journal, 33(1), 415-422. https://doi.org/10.57239/prn.25.03310048

10. Raihena, S. S., Arif, M. A. H., Mariam, Q. R., Hussain, A. H., & Rashaq, M. AI-Enhanced Decision Support Systems for Autism Caregivers: Redefining HR's Role in Workforce Planning and Patient-Centered Care. https://doi.org/10.63682/fhi2698