RESEARCH ARTICLE OPEN ACCESS
https://www.ijcsejournal.org/ ***Type of Article*** *(Review Article)*

Explainable Artificial Intelligence for Adaptive Virtual Reality Systems: A Conceptual Framework for Transparent User Adaptation

Shivani
Email: rshivu93@gmail.com

Abstract

Adaptive Virtual Reality (VR) systems are increasingly incorporating artificial intelligence (AI) to personalize user experiences through continuous adjustments to task difficulty, navigation support, interaction strategies, comfort settings, and narrative progression. By analyzing multimodal data such as gaze behaviour, physiological responses, movement patterns, and task performance, these systems can dynamically modify virtual environments to enhance engagement and performance. Despite these advances, the decision-making processes underlying AI-driven adaptations remain largely opaque, limiting users' understanding of why specific adaptations occur and potentially reducing trust, perceived fairness, and technology acceptance. While Explainable Artificial Intelligence (XAI) has emerged as an effective approach for improving transparency in AI-enabled decision-making, most existing XAI techniques have been developed for discrete prediction tasks and are not readily applicable to the continuous, immersive, and context-sensitive nature of adaptive VR environments.

*This paper proposes a **seven-layer conceptual framework** for integrating explainability into adaptive VR systems while preserving user immersion and interaction quality. The framework consists of seven interconnected components: **Adaptation Trigger Layer, Inference and Adaptation Engine, Explanation Generation Layer, Explanation Modality Layer, Granularity and Timing Control, Trust Calibration Loop, and Evaluation Layer**. Unlike existing approaches that primarily emphasize personalization accuracy, the proposed framework treats explainability as a core design principle and explicitly addresses the trade-offs between transparency, immersion, usability, and user trust. In addition, the framework incorporates ethical considerations related to privacy, physiological data processing, and responsible AI, providing a broader human-centred perspective for future adaptive VR systems. The paper concludes by outlining future research directions and offering practical guidance for researchers and designers seeking to develop transparent, trustworthy, and user-centred adaptive virtual reality applications.*

Keywords: *Explainable Artificial Intelligence (XAI), Adaptive Virtual Reality, Human-Centred AI, Trustworthy AI, Transparency, User Trust, Immersive Systems, Human–Computer Interaction.*

Publisher
https://www.ijcsejournal.org/ **Type of Article** *(Review Article)*

1. Introduction

Virtual Reality (VR) has evolved from delivering static, computer-generated environments into intelligent immersive systems capable of continuously adapting to users' behaviours, preferences, and contextual conditions. Recent advances in artificial intelligence (AI), machine learning (ML), multimodal sensing, and generative AI have accelerated this transformation, enabling VR systems to dynamically personalize user experiences across domains such as education, healthcare, rehabilitation, industrial training, entertainment, and collaborative virtual environments (Rahimi et al., 2025; Wang et al., 2025; Baker et al., 2022). Adaptive VR systems continuously monitor behavioural, physiological, and contextual information—including gaze behaviour, locomotion, task performance, head and hand movements, heart rate, and electrodermal activity—to estimate users' cognitive, emotional, and physical states, allowing virtual environments to modify task difficulty, navigation assistance, interaction strategies, and narrative progression in real time (Li et al., 2025; Martin et al., 2020; Kritikos et al., 2021). These adaptive capabilities have significantly enhanced personalization, user engagement, and task effectiveness while expanding the practical applicability of immersive technologies.

The growing integration of AI into VR has shifted adaptive mechanisms from rule-based systems toward intelligent data-driven decision-making models capable of learning from continuous user interaction. Modern adaptive VR platforms increasingly employ supervised learning, reinforcement learning, deep learning, and large language models (LLMs) to infer user intent and optimize immersive experiences through closed-loop adaptation (Li et al., 2025; Rahimi et al., 2025). These AI-driven approaches enable systems to predict user needs, personalize learning pathways, mitigate cybersickness, improve therapeutic interventions, and support adaptive navigation without requiring explicit user input (Martin et al., 2020; Besga et al., 2025). Despite these advances, increasing algorithmic complexity has reduced the transparency of adaptation decisions, making it difficult for users to understand why the system modifies its behaviour during immersive interaction.

The lack of transparency has emerged as one of the principal barriers to trustworthy AI-enabled immersive systems. Users often observe changes in task difficulty, environmental conditions, interaction styles, or feedback mechanisms without understanding the reasoning behind these adaptations. Such opacity may reduce user trust, perceived fairness, technology acceptance, and willingness to rely on adaptive systems, particularly when adaptation decisions are derived from implicit physiological or behavioural signals that users cannot directly observe or verify (Guidotti et al., 2018; Arrieta et al., 2020; Gunning et al., 2019). These challenges are amplified in VR because AI decisions influence not merely a prediction or recommendation but the user's ongoing perceptual and experiential reality. Consequently, transparency becomes not only a technical requirement but also a human-centred design principle essential for trustworthy immersive interaction.

Explainable Artificial Intelligence (XAI) has emerged as a promising research paradigm for improving the interpretability and transparency of complex AI models by providing human-understandable explanations for automated decisions. Existing XAI techniques—including Local Interpretable Model-agnostic Explanations (LIME), SHapley Additive exPlanations (SHAP), counterfactual reasoning, surrogate modelling, and attention-based visualization—have demonstrated considerable success across domains such as healthcare, finance, autonomous systems, and recommender systems (Ribeiro et al., 2016; Lundberg & Lee, 2017; Arrieta et al., 2020). More recent research has further emphasized adaptive and human-centred explanations that consider users' expertise, cognitive characteristics, contextual needs, and explanation preferences rather than providing identical explanations to all users (Turchi et al., 2025).

Although both adaptive VR and XAI have advanced considerably during the last decade, their integration remains limited. Current adaptive VR research primarily emphasizes prediction accuracy, personalization effectiveness, responsiveness, and user performance, whereas XAI research has predominantly focused on conventional machine learning tasks involving isolated predictions rather than continuously evolving immersive environments (Li et al., 2025; Rahimi et al., 2025). Recent systematic reviews of adaptive VR have highlighted the growing adoption of AI-

Publisher
https://www.ijcsejournal.org/ **Type of Article** *(Review Article)*

driven personalization while simultaneously identifying transparency, explainability, privacy, and user trust as important unresolved research challenges (Correia, 2024; Wang et al., 2025). Similarly, recent surveys on Explainable AI in Extended Reality conclude that existing XAI techniques require substantial adaptation before they can effectively support immersive environments where explanations must preserve presence, minimize cognitive disruption, and respect user experience constraints.

The challenge extends beyond explaining individual AI decisions. Adaptive VR systems operate as continuous closed-loop environments in which sensing, inference, adaptation, and user response occur simultaneously. Explanations delivered at inappropriate moments may interrupt immersion, increase cognitive workload, or diminish users' sense of presence. Consequently, explanation design within immersive environments must carefully balance transparency with usability, responsiveness, and uninterrupted interaction. Existing literature provides valuable insights into explainability and adaptive personalization independently; however, a comprehensive framework that systematically integrates explainability throughout the adaptive VR lifecycle—from adaptation triggers and AI inference to explanation delivery, trust calibration, ethical governance, and evaluation—remains largely absent.

To address this research gap, this paper proposes a **seven-layer conceptual framework** for integrating Explainable Artificial Intelligence into adaptive Virtual Reality systems. Unlike existing approaches that treat explainability as an auxiliary interface feature, the proposed framework positions transparency as a core architectural principle embedded throughout the adaptive decision-making process. The framework comprises seven interrelated layers: **Adaptation Trigger Layer, Inference and Adaptation Engine, Explanation Generation Layer, Explanation Modality Layer, Granularity and Timing Control, Trust Calibration Loop, and Evaluation Layer**. Together, these layers support the design of adaptive VR systems that are not only intelligent and personalized but also transparent, trustworthy, ethically responsible, and human-centred.

The primary contributions of this paper are fourfold. First, it synthesizes current research at the intersection of Explainable Artificial Intelligence and Adaptive Virtual Reality while identifying critical research gaps that limit transparent immersive systems. Second, it proposes a comprehensive seven-layer conceptual framework that embeds explainability throughout the adaptive VR lifecycle. Third, it extends existing conceptualizations by integrating trust calibration, privacy, ethical AI, and human-centred explanation design into immersive adaptive systems. Finally, it outlines future research directions and provides a foundation for the empirical evaluation of transparent, trustworthy, and explainable adaptive Virtual Reality systems.

2. Research Gaps and Problem Statement

Although Explainable Artificial Intelligence (XAI) and Adaptive Virtual Reality (VR) have both experienced significant advances over the past decade, research at their intersection remains fragmented. XAI has primarily focused on improving the transparency and interpretability of complex machine learning models across domains such as healthcare, finance, autonomous systems, and decision support, where AI outputs are generally discrete and evaluated independently of the surrounding interaction context (Guidotti et al., 2018; Arrieta et al., 2020; Gunning et al., 2019). In contrast, adaptive VR research has concentrated on optimizing personalization, user engagement, and system responsiveness through real-time behavioural, contextual, and physiological adaptation, with comparatively limited attention given to explaining why adaptive decisions occur or how those explanations should be communicated within immersive environments (Kritikos et al., 2021; Martin et al., 2020).

Recent reviews further highlight this disconnect. Li et al. (2025), in a comprehensive review of personalized VR systems, report increasing adoption of machine learning and multimodal sensing for adaptive interaction while observing that explainability remains largely peripheral to system design. Although techniques such as SHAP and feature attribution are beginning to appear within adaptive VR research, most personalization models continue to operate as opaque black boxes, particularly when deep learning, reinforcement learning, or generative AI

Publisher
https://www.ijcsejournal.org/ **Type of Article** (Review Article)

components are incorporated. Similarly, Rahimi et al. (2025) identify explainability as one of the major unresolved challenges in AI-driven VR, emphasizing that future immersive systems must provide transparent and human-understandable adaptation mechanisms without compromising real-time performance or user immersion.

Beyond these observations, several fundamental research limitations remain unresolved. First, most XAI techniques—including LIME, SHAP, and counterfactual explanations—were originally developed for static prediction tasks and therefore provide limited guidance for continuously evolving adaptive environments in which decisions are made repeatedly throughout user interaction (Ribeiro et al., 2016; Lundberg & Lee, 2017). Second, existing adaptive VR systems predominantly evaluate personalization in terms of prediction accuracy, task performance, or user engagement, whereas transparency, explanation quality, and calibrated trust are rarely incorporated as primary design objectives (Cancro et al., 2022; Newen et al., 2025). Third, adaptation decisions frequently rely on implicit physiological signals such as gaze behaviour, heart rate, electrodermal activity, and body movement, yet users are seldom informed how these signals influence AI-driven adaptations, creating concerns regarding transparency, informed consent, privacy, and algorithmic fairness (Martin et al., 2020; Kritikos et al., 2021).

An additional limitation concerns evaluation methodology. Existing XAI research typically assesses explanation quality using metrics such as fidelity, interpretability, and comprehensibility, whereas VR research focuses on presence, immersion, cybersickness, and task performance. Few studies integrate these complementary perspectives into a unified evaluation strategy capable of assessing both explanation effectiveness and immersive user experience simultaneously (Sameri et al., 2024). Consequently, there is currently no widely accepted framework for evaluating explainable adaptive VR systems from technical, human-centred, and ethical perspectives.

Collectively, these limitations reveal that existing research has addressed explainability and adaptive VR as largely independent problems. While previous studies recommend increasing transparency within intelligent VR systems, they provide limited guidance regarding **where explanations should be generated, how they should be communicated, when they should be presented, how they should adapt to individual users, and how their effectiveness should be evaluated without disrupting immersion**. This absence of a comprehensive design methodology represents a significant research gap.

To address these limitations, this study proposes a **seven-layer conceptual framework** that systematically embeds explainability throughout the adaptive VR lifecycle. Rather than viewing explanation as an additional interface component introduced after adaptive decisions have been made, the proposed framework positions transparency as a fundamental architectural principle governing adaptation triggers, AI inference, explanation generation, explanation delivery, trust calibration, ethical governance, and multidimensional evaluation. By integrating concepts from Explainable AI, Human–Computer Interaction, Trustworthy AI, and Adaptive Virtual Reality, the framework provides a foundation for designing immersive systems that are simultaneously intelligent, transparent, trustworthy, and human-centred.

Table 1. Research gaps motivating the proposed framework

Identified Gap	Description	Supporting Sources
Discrete-decision bias in XAI methods	Established XAI techniques such as SHAP and LIME were developed for single, discrete predictions and do not map cleanly onto continuously evolving adaptation loops typical of VR systems.	Ribeiro et al. (2016); Lundberg & Lee (2017)
Immersion-transparency tradeoff is under-specified	Prior work acknowledges that explanation delivery can disrupt presence, but few studies offer design guidance on when and how to explain without breaking immersion.	Rahimi et al. (2025); Li et al. (2025)

Publisher
https://www.ijcsejournal.org/ **Type of Article** *(Review Article)*

Identified Gap	Description	Supporting Sources
Physiological inference is rarely explained to users	Systems that adapt based on heart rate, gaze, or electrodermal activity seldom disclose to users what these signals are inferred to mean, raising both comprehension and consent concerns.	Kritikos et al. (2021); Martin et al. (2020)
Lack of VR-specific evaluation standards for explanations	Most XAI evaluation relies on fidelity and comprehensibility metrics from non-immersive contexts; VR-specific companions such as presence and cybersickness are rarely combined with explanation quality metrics.	Sameri et al. (2024)
Absence of unified conceptual frameworks	Explainability and adaptive VR personalization have largely been studied as separate literatures, with limited integration into a single design framework that spans triggers, inference, explanation, and evaluation.	Li et al. (2025); Rahimi et al. (2025)

Taken together, these gaps motivate a framework-level contribution rather than a narrow technical one: what is needed is not simply a new attribution algorithm, but a structured account of where explainability must be embedded across the pipeline of an adaptive VR system, and how explanation delivery should be shaped by the constraints of immersive, embodied interaction.

3. Literature Review

3.1 Explainable AI: Concepts, Methods, and Human-Centred Perspectives

The increasing deployment of complex machine learning models across high-impact domains has intensified concerns regarding transparency, accountability, and user trust, leading to the rapid evolution of Explainable Artificial Intelligence (XAI) (Gunning et al., 2019; Arrieta et al., 2020). XAI aims to make AI-driven decisions understandable by providing explanations that enable users to interpret model behaviour and evaluate system reliability. Existing approaches are commonly categorized into model-specific and model-agnostic methods, with explanations generated at either global or local levels depending on whether the objective is to explain overall model behaviour or individual predictions (Guidotti et al., 2018).

Among model-agnostic techniques, **LIME** and **SHAP** have emerged as the most widely adopted explanation methods because they can interpret predictions without requiring access to a model's internal architecture. LIME generates local surrogate models to approximate decision boundaries, whereas SHAP employs cooperative game theory to quantify the contribution of individual features to model predictions (Ribeiro et al., 2016; Lundberg & Lee, 2017). Although SHAP offers stronger theoretical guarantees and supports both local and global interpretability, its computational complexity limits real-time deployment. Conversely, LIME provides greater computational efficiency but may produce unstable explanations across repeated executions. Counterfactual explanations further extend XAI by identifying minimal input changes required to alter a model's output, offering explanations that closely align with human reasoning despite ongoing challenges related to feasibility and consistency (Arrieta et al., 2020).

Recent research has shifted the focus of XAI from algorithmic interpretability toward **human-centred explainability**, emphasizing that effective explanations should improve user understanding, calibrated trust, and decision quality rather than simply exposing model internals (Newen et al., 2025). Studies on intelligibility further demonstrate that explanations describing why a system acted in a particular manner are generally more useful than explanations describing alternative outcomes (Lim & Dey, 2009). Moreover, explanation requirements vary across users and contexts, suggesting that explanation strategies should account for user expertise, goals, and social context rather than relying on one-size-fits-all approaches (Ehsan et al., 2021). Despite these advances, most XAI techniques have been developed for static prediction tasks and provide limited guidance for continuously adaptive immersive environments.

Publisher
https://www.ijcsejournal.org/ **Type of Article** *(Review Article)*

3.2 Adaptive and Personalized Virtual Reality Systems

Adaptive Virtual Reality has evolved from rule-based interaction models toward AI-driven systems capable of dynamically personalizing immersive experiences using behavioural, physiological, and contextual information. Recent advances in machine learning and multimodal sensing enable VR systems to continuously estimate user states from gaze behaviour, movement patterns, task performance, and physiological signals, allowing environments to modify navigation support, interaction strategies, task difficulty, and therapeutic interventions in real time (Li et al., 2025).

Current adaptive VR approaches can generally be categorized as **reactive** or **proactive**. Reactive systems respond after changes in user behaviour are detected, whereas proactive systems attempt to anticipate future user states before they become observable (Li et al., 2025). Although proactive adaptation offers greater personalization potential, computational complexity, prediction uncertainty, and real-time performance constraints have limited its widespread implementation. Consequently, many contemporary adaptive VR systems employ hybrid architectures that combine reactive monitoring with higher-level AI-driven personalization.

Physiological sensing has become particularly important in adaptive VR applications involving user comfort and healthcare. Studies have demonstrated that signals such as heart rate and electrodermal activity can accurately predict cybersickness, enabling systems to adjust virtual environments before discomfort becomes severe (Martin et al., 2020). Similar adaptive mechanisms have been successfully applied in virtual exposure therapy, where physiological responses guide the dynamic adjustment of therapeutic stimuli (Kritikos et al., 2021). Despite these advances, adaptation decisions are typically generated by opaque AI models, leaving users unaware of why particular modifications occur during immersive interaction. This lack of transparency remains one of the principal limitations of current adaptive VR systems.

3.3 Explainable AI in Virtual Reality

Research integrating XAI with Virtual Reality remains relatively limited but has grown steadily in recent years. Existing studies primarily apply explainability techniques to physiological prediction models, where feature-attribution methods are used to identify the behavioural or physiological variables responsible for predictions such as cybersickness or user engagement (Sameri et al., 2024). Other work argues that explainability will become increasingly important as generative AI, large language models, and multimodal interaction are integrated into immersive environments, requiring users to understand AI-generated adaptations without disrupting presence or interaction quality (Rahimi et al., 2025).

Although these studies demonstrate the feasibility of applying XAI within immersive environments, they remain narrowly focused on specific prediction tasks. Current research provides limited guidance regarding how explanations should be generated, presented, adapted to different users, or integrated into the overall architecture of adaptive VR systems. Consequently, explainability is often treated as an additional interface feature rather than as a fundamental design principle embedded throughout the adaptive decision-making process.

3.4 Trust Calibration and Evaluation

Trust is a critical determinant of user acceptance in adaptive VR because AI-driven adaptations continuously influence the immersive experience while remaining largely invisible to users. Trust

Publisher
https://www.ijcsejournal.org/ **Type of Article** (Review Article)

calibration research emphasizes that explanations should enable users to develop confidence that accurately reflects system reliability rather than encouraging either over-trust or unnecessary skepticism (Cancro et al., 2022; Newen et al., 2025). Effective explanations therefore depend not only on content but also on timing, presentation modality, interactivity, and user context.

Evaluation of explainable adaptive VR systems requires integrating conventional XAI metrics with measures commonly used in immersive computing. Existing XAI research emphasizes explanation fidelity, interpretability, and comprehensibility, whereas VR research evaluates presence, immersion, cybersickness, cognitive workload, and user performance (Martin et al., 2020; Sameri et al., 2024). Recent studies further highlight the importance of user trust, perceived fairness, transparency, privacy awareness, and technology acceptance as complementary evaluation dimensions for human-centred AI systems. A comprehensive evaluation framework should therefore assess both technical explanation quality and the broader human experience to ensure that explainability enhances rather than disrupts immersive interaction.

Literature Synthesis

The reviewed literature demonstrates significant advances in both Explainable Artificial Intelligence and Adaptive Virtual Reality; however, these research streams have progressed largely independently. Existing XAI methods have been primarily designed for static machine learning applications, whereas adaptive VR research has prioritized personalization accuracy, responsiveness, and user engagement with comparatively limited attention to transparent decision-making. Furthermore, current studies do not provide a comprehensive framework that systematically integrates adaptation triggers, AI inference, explanation generation, explanation delivery, trust calibration, ethical considerations, and multidimensional evaluation within immersive environments. These limitations establish the theoretical foundation for the seven-layer conceptual framework proposed in the following section, which seeks to embed explainability as a core architectural principle for transparent, trustworthy, and human-centred adaptive Virtual Reality systems.

4. Proposed Conceptual Framework

Building on the gaps and literature reviewed above, this section presents a seven-layer conceptual framework, summarized in Table 2, for embedding explainability into adaptive VR systems. The framework is organized as a pipeline: signals flow from the adaptation trigger layer through an inference engine into explanation generation, are shaped by the explanation modality and timing layers before reaching the user, and are continuously assessed through a trust calibration loop and an evaluation layer that in turn informs redesign of the earlier layers.

Table 2. The seven-layer XAI-for-adaptive-VR framework

Framework Layer	Core Function	Representative Techniques / Design Choices
Adaptation Trigger Layer	Captures the signals that drive real-time adaptation (behavioral, physiological, contextual)	Gaze and attention tracking; performance/error logs; heart-rate and galvanic skin response; locomotion and posture data
Inference / Adaptation Engine	Maps triggers to adaptation decisions using ML or hybrid rule-based models	Reinforcement learning policies; supervised classifiers for state estimation; LLM-driven narrative or dialogue agents
Explanation Generation Layer	Produces a human-interpretable account of why an adaptation occurred	Local surrogate models (LIME); Shapley-value attributions (SHAP); counterfactual reasoning; rule extraction

Publisher
https://www.ijcsejournal.org/ Type of Article (Review Article)

Framework Layer	Core Function	Representative Techniques / Design Choices
Explanation Modality Layer	Delivers the explanation through a channel appropriate to the immersive context	Diegetic in-world cues; non-diegetic overlays; post-session debrief; ambient/peripheral signals
Granularity and Timing Control	Adjusts explanation depth and delivery moment to user expertise and task state	User-set verbosity sliders; context-sensitive deferral; flow-state detection to suppress interruption
Trust Calibration Loop	Feeds explanation outcomes back into the system to align user trust with actual reliability	Confidence scoring; explanation-seeking interaction logging; longitudinal trust surveys
Evaluation Layer	Measures whether explanations achieve transparency without harming the VR-specific experience	Presence questionnaires (IPQ); Simulator Sickness Questionnaire; explanation satisfaction scales; task performance metrics

4.1 Adaptation Trigger Layer

The **Adaptation Trigger Layer** serves as the foundation of the proposed framework by defining the categories of information that may legitimately initiate adaptive behaviour within a Virtual Reality (VR) environment. The primary objective of this layer is to collect and integrate multimodal data that accurately reflects the user's current cognitive, behavioural, physiological, and contextual state, thereby enabling timely and personalized system adaptations. These adaptation triggers may originate from **explicit user inputs**, such as task performance, interaction history, navigation choices, and stated preferences, as well as **implicit indicators**, including gaze behaviour, locomotion patterns, heart rate, electrodermal activity, head movements, and environmental context (Alghofaili et al., 2019; Martin et al., 2020). By combining multiple information sources, the system can construct a more reliable representation of the user's state than would be possible through a single modality alone.

Unlike conventional adaptive VR architectures that primarily emphasize prediction accuracy, the proposed framework introduces **transparency at the point of data acquisition**. Before adaptive decisions are generated, users should be informed about the categories of information being collected, the purpose of data collection, and how these signals contribute to subsequent adaptations. This recommendation is particularly important for physiological data because users are often unaware that signals such as heart rate or skin conductance can influence system behaviour. Providing clear onboarding explanations and obtaining informed consent enhance transparency while preserving immersion by avoiding unnecessary interruptions during real-time interaction. This approach aligns with the ethical principles of trustworthy AI, including user autonomy, informed consent, and responsible data governance.

Another important design consideration concerns the **quality and reliability of adaptation triggers**. Physiological and behavioural signals are inherently noisy and may be affected by sensor inaccuracies, environmental conditions, or individual differences. Consequently, the framework recommends adopting a **multimodal sensing strategy**, whereby multiple behavioural and physiological indicators are analysed together before initiating an adaptive response. Such an approach reduces false adaptation decisions and increases the robustness of user-state estimation, particularly in applications involving healthcare, rehabilitation, education, and industrial training.

The proposed framework also recognizes that not all adaptation triggers require the same level of transparency. Routine adaptations, such as minor interface adjustments or pacing modifications, may only require general disclosure during system onboarding. In contrast, adaptations that substantially influence user experience, learning progression, or therapeutic intervention should provide users with accessible explanations regarding the primary

Publisher
https://www.ijcsejournal.org/ **Type of Article** (Review Article)

factors that influenced the decision. This selective transparency balances user understanding with the need to preserve immersion and cognitive flow.

Overall, the Adaptation Trigger Layer establishes the foundation for trustworthy adaptive VR by ensuring that adaptive decisions originate from reliable, ethically collected, and transparently communicated user information. By embedding transparency and responsible data collection at the earliest stage of the adaptation pipeline, the framework supports explainability throughout the subsequent inference, explanation generation, and trust calibration processes.

4.2 Inference and Adaptation Engine

The **Inference and Adaptation Engine** constitutes the computational intelligence of the proposed framework, transforming multimodal user information collected by the Adaptation Trigger Layer into adaptive decisions that personalize the Virtual Reality (VR) experience. This layer integrates behavioural, physiological, and contextual signals to infer the user's current state—such as cognitive workload, emotional engagement, learning progress, navigation efficiency, or discomfort—and subsequently determines the most appropriate adaptation strategy. Depending on the application domain, these adaptations may include modifying task difficulty, providing navigational assistance, adjusting environmental conditions, personalizing instructional content, or altering narrative progression (Li et al., 2025; Rahimi et al., 2025).

The proposed framework is intentionally **model-agnostic**, allowing developers to employ a wide range of computational approaches, including supervised machine learning, reinforcement learning, Bayesian inference, hybrid rule-based systems, and Large Language Model (LLM)-based intelligent agents. This flexibility ensures that the framework remains applicable across diverse adaptive VR applications, including education, healthcare, rehabilitation, industrial training, and immersive entertainment. Rather than prescribing a particular algorithm, the framework emphasizes that the selected inference model should achieve an appropriate balance between predictive accuracy, computational efficiency, and interpretability.

A distinguishing characteristic of this framework is that **explainability is considered a design requirement rather than an afterthought**. Consequently, every inference model should expose an interpretable interface capable of supporting the Explanation Generation Layer. Depending on the underlying algorithm, this interface may include feature importance scores, attention weights, confidence estimates, decision rules, latent-state representations, or automatically generated natural-language rationales. By making intermediate reasoning accessible, the framework enables explanation techniques to communicate not only *what* adaptation occurred but also *why* the adaptation was selected.

The framework further recognizes that highly sophisticated AI models, particularly deep neural networks, reinforcement learning policies, and generative AI systems, often provide superior personalization while simultaneously reducing transparency. To address this challenge, the framework recommends **hybrid inference architectures** that combine high-performance adaptive models with complementary interpretable components. For example, an LLM-driven narrative agent may generate personalized dialogue or learning scenarios, while a lightweight supervised classifier or rule-based module provides an interpretable representation of the key behavioural or physiological factors influencing the adaptation. Such hybrid architectures improve explanation fidelity while preserving the flexibility and responsiveness required for real-time immersive environments, consistent with emerging adaptive VR architectures reported by Li et al. (2025).

From a system-design perspective, the Inference and Adaptation Engine should also support **continuous learning and dynamic model refinement**. User behaviour and preferences frequently evolve over repeated VR sessions; therefore, adaptive models should update their internal representations while maintaining explanation consistency and avoiding unpredictable behavioural changes that could undermine user trust. Confidence estimation and

Publisher
https://www.ijcsejournal.org/ Type of Article (Review Article)

uncertainty awareness should likewise accompany adaptive decisions, allowing subsequent layers to determine whether explanations require additional detail or whether user confirmation should be requested before implementing high-impact adaptations.

Overall, the Inference and Adaptation Engine serves as the central decision-making component of the proposed framework, bridging multimodal user sensing with transparent adaptive behaviour. By combining computational intelligence with interpretable decision interfaces, the layer establishes the foundation upon which meaningful explanations, calibrated trust, and responsible adaptive Virtual Reality experiences can be built.

4.3 Explanation Generation Layer

The **Explanation Generation Layer** transforms the outputs of the Inference and Adaptation Engine into explanations that enable users to understand the reasoning behind adaptive decisions. Acting as the central explainability component of the proposed framework, this layer bridges the gap between complex AI decision-making and human interpretation by converting computational evidence into explanations that are meaningful, concise, and relevant to the user's current context. Rather than exposing internal model complexity, the objective is to communicate **why** a particular adaptation occurred, **which factors** influenced the decision, and **how** alternative user behaviours or contextual conditions could have produced different outcomes.

The framework adopts a **method-agnostic approach** to explanation generation, allowing different XAI techniques to be selected according to the characteristics of the underlying AI model, computational requirements, and the importance of the adaptation. Local explanation methods such as **Local Interpretable Model-agnostic Explanations (LIME)** are particularly suitable for explaining individual adaptation events because they approximate the behaviour of complex models around a specific decision (Ribeiro et al., 2016). In contrast, **SHapley Additive exPlanations (SHAP)** provide feature-attribution scores that identify the relative contribution of different behavioural, physiological, or contextual variables to adaptive decisions over a broader interaction period, making them appropriate for analysing long-term personalization patterns and user-state estimation (Lundberg & Lee, 2017; Sameri et al., 2024).

For adaptations that substantially influence user experience, such as changes in task difficulty, learning pathways, therapeutic intensity, or navigation assistance, the framework recommends **counterfactual explanations**. These explanations describe the minimal changes in user behaviour or contextual conditions that would have resulted in a different adaptive outcome, thereby supporting actionable understanding and encouraging user reflection. Previous research has shown that explanations framed around *why a system acted* and *what users could do differently* are generally easier to understand and more useful than explanations that merely describe system behaviour without practical guidance (Lim & Dey, 2009). Consequently, counterfactual reasoning is expected to improve user comprehension while promoting greater engagement with adaptive VR systems.

The proposed framework further recognizes that explanation quality extends beyond algorithmic accuracy. Effective explanations should be **accurate, faithful to the underlying AI model, concise, context-aware, and appropriate for the user's expertise**. Novice users may require simple natural-language explanations that emphasise the primary factors influencing an adaptation, whereas experienced users or domain experts may benefit from more detailed information, including confidence scores, feature contributions, or model uncertainty estimates. Accordingly, explanation generation should be viewed as a dynamic and user-centred process rather than a static computational output.

To support diverse adaptive VR applications, the framework recommends selecting explanation techniques according to the **impact and complexity of the adaptive decision**. Lightweight local explanations are appropriate for frequent, low-risk adaptations where computational efficiency is essential, whereas attribution-based or counterfactual explanations are more suitable for high-impact decisions that significantly influence user

Publisher
https://www.ijcsejournal.org/ Type of Article (Review Article)

performance, safety, or therapeutic outcomes. This adaptive explanation strategy enables the framework to balance computational efficiency, interpretability, and user understanding across different application domains.

Overall, the Explanation Generation Layer establishes the foundation for transparent adaptive behaviour by converting complex AI reasoning into understandable and actionable explanations. By integrating multiple explanation strategies while remaining independent of any specific AI algorithm, this layer ensures that subsequent explanation delivery mechanisms can communicate adaptive decisions in a manner that promotes transparency, user trust, and informed interaction without compromising the intelligence of the underlying adaptive system.

4.4 Explanation Modality Layer

The **Explanation Modality Layer** determines **how** explanations generated by the previous layer are communicated to users within the Virtual Reality (VR) environment. Unlike conventional Explainable Artificial Intelligence (XAI) systems, where explanations are typically presented as textual descriptions or graphical visualizations on two-dimensional interfaces, adaptive VR requires explanation delivery mechanisms that preserve immersion while maintaining sufficient transparency. Consequently, this layer represents one of the principal innovations of the proposed framework by addressing the fundamental trade-off between **user understanding** and **immersive experience** identified in recent adaptive VR research (Li et al., 2025; Rahimi et al., 2025).

The framework proposes four complementary explanation modalities that can be selected according to the context, significance, and urgency of the adaptive decision. **Diegetic explanations** are embedded directly within the virtual environment and delivered through narratively meaningful elements, such as comments from non-player characters (NPCs), environmental changes, interactive virtual objects, or contextual visual cues. Because these explanations become part of the virtual world's narrative, they preserve the user's sense of presence while providing understandable justification for adaptive system behaviour. Such explanations are particularly appropriate for educational simulations, serious games, and narrative-driven VR experiences.

In situations where precision and immediate user awareness are more important than uninterrupted immersion, the framework recommends **non-diegetic explanations**. These explanations are presented through interface elements external to the virtual narrative, including heads-up displays (HUDs), pop-up notifications, icons, or textual overlays. Although these approaches may temporarily reduce immersion, they are suitable for high-priority adaptations involving user safety, cybersickness mitigation, therapeutic interventions, or critical task guidance, where clear communication is essential.

For adaptive decisions that do not require immediate explanation, the framework introduces **post-hoc debriefing** as an alternative modality. Rather than interrupting ongoing interaction, explanations are deferred until a natural pause in the session or presented after the immersive experience has concluded. Post-session summaries can provide users with a comprehensive overview of adaptive decisions, behavioural trends, and system reasoning while avoiding unnecessary disruption during real-time interaction. This modality is particularly valuable in educational, rehabilitation, and professional training environments where reflective learning and performance review are important objectives.

The framework also incorporates **ambient or peripheral explanations**, which provide subtle indications that an adaptive event has occurred without delivering a detailed justification. Examples include gentle audio feedback, changes in environmental lighting, haptic cues, or minor interface animations that communicate adaptive system activity while imposing minimal cognitive load. These lightweight explanations are well suited to frequent, low-impact adaptations where continuous detailed explanations would become distracting or contribute to explanation fatigue.

Publisher
https://www.ijcsejournal.org/ Type of Article (Review Article)

Rather than prescribing a single explanation modality, the proposed framework recommends **context-aware modality selection** based on the **salience, urgency, and potential impact** of the adaptive decision. Routine, low-risk adaptations—such as pacing adjustments or minor interface refinements—can be communicated effectively through ambient signals or diegetic cues that preserve immersion. In contrast, high-impact adaptations affecting user safety, therapeutic interventions, learning progression, or significant task difficulty should employ non-diegetic or post-hoc explanations that provide greater detail and justification. This adaptive selection strategy enables the framework to balance transparency with uninterrupted user experience while minimizing cognitive disruption.

The framework further recognizes that explanation modality should be **personalized** according to user preferences, expertise, accessibility requirements, and application context. Novice users may benefit from richer visual or textual explanations, whereas experienced users may prefer concise notifications or optional explanations that can be accessed on demand. Supporting multiple modalities also improves accessibility for users with diverse perceptual or cognitive needs and enables adaptive VR systems to accommodate a wider range of interaction scenarios.

4.5 Granularity and Timing Control

The **Granularity and Timing Control** layer governs **how much explanatory information is provided** and **when it should be delivered** during adaptive Virtual Reality (VR) interactions. While the Explanation Generation Layer determines the content of an explanation and the Explanation Modality Layer specifies its delivery channel, this layer ensures that explanations are presented at an appropriate level of detail and at moments that maximize user understanding while minimizing disruption to immersive experiences. Its primary objective is to balance transparency with cognitive continuity, recognizing that excessive or poorly timed explanations may interrupt user engagement, increase cognitive workload, and reduce the sense of presence.

The proposed framework adopts a **user-centred and adaptive approach** to explanation granularity. Rather than assuming identical information requirements for all users, the framework recommends that explanation depth be dynamically adjusted according to factors such as user expertise, task complexity, interaction history, and individual preferences. Novice users may benefit from detailed explanations that describe the reasoning behind adaptive decisions, whereas experienced users often require only concise notifications or brief justifications. To accommodate these differences, explanation granularity should be configurable through user-selectable preference settings, ranging from **minimal explanations**, consisting of subtle acknowledgement signals, to **comprehensive explanations** that include feature contributions, confidence estimates, and counterfactual reasoning.

Equally important is the **timing of explanation delivery**. Unlike conventional XAI systems, where explanations are typically presented immediately after an AI decision, adaptive VR environments require timing strategies that preserve immersion and support uninterrupted interaction. Consistent with the trust calibration framework proposed by Cancro et al. (2022), explanation timing should be regarded as a dynamic design parameter rather than a fixed property of the explanation itself. Immediate explanations are appropriate when adaptive decisions influence user safety, therapeutic outcomes, or critical learning objectives, whereas routine personalization decisions can often be deferred until natural interaction pauses or presented only when requested by the user.

To facilitate intelligent timing decisions, the framework recommends monitoring behavioural indicators that reflect the user's current level of engagement or cognitive demand. Measures such as sustained task performance, stable gaze behaviour, interaction frequency, or estimated flow state can be used to identify periods during which interruptions are likely to reduce immersion. During such high-engagement intervals, the system should prioritise concise or deferred explanations, reserving more detailed explanations for moments of lower cognitive load or post-task reflection. This adaptive timing strategy reduces unnecessary interruptions while ensuring that important adaptive decisions remain transparent and understandable.

Publisher
https://www.ijcsejournal.org/ **Type of Article** (Review Article)

The framework further recognizes that both explanation granularity and timing should evolve throughout long-term interaction. As users become more familiar with system behaviour, their explanatory requirements are likely to change. Consequently, adaptive VR systems should continuously learn from user feedback, explanation requests, and interaction patterns to refine future explanation strategies. For example, users who consistently request additional details may automatically receive richer explanations, whereas users who frequently dismiss explanations may benefit from more concise communication. This continuous personalization supports efficient human–AI interaction while avoiding explanation fatigue.

4.6 Trust Calibration Loop

The **Trust Calibration Loop** represents the adaptive feedback mechanism of the proposed framework, ensuring that user trust evolves in accordance with the actual reliability and performance of the adaptive Virtual Reality (VR) system. Unlike conventional explainable AI approaches that treat explanations as one-way disclosures, the proposed framework views explanation as an **interactive and continuous process** in which user responses influence future explanation strategies. The primary objective of this layer is not to maximize user trust unconditionally, but to establish **appropriately calibrated trust**, whereby users rely on the system when its recommendations are reliable while maintaining healthy skepticism when uncertainty or potential errors are present (Cancro et al., 2022; Newen et al., 2025).

To achieve this objective, the framework continuously monitors both **subjective** and **behavioural** indicators of user trust. Subjective indicators include explicit feedback such as explanation requests, user satisfaction ratings, trust questionnaires, and the acceptance or dismissal of explanations. Behavioural indicators include interaction patterns such as compliance with system recommendations, frequency of manual overrides, response times, adaptation acceptance rates, repeated explanation requests, and navigation or task-performance behaviour following adaptive interventions. Collectively, these measures provide a comprehensive understanding of how users perceive and respond to adaptive system behaviour beyond traditional self-reported trust measures.

The framework recommends using these trust indicators to **dynamically personalize future explanations**. For example, users who frequently request additional information or hesitate before accepting adaptive recommendations may benefit from more detailed explanations, richer visualizations, or increased transparency regarding model confidence. Conversely, users who consistently demonstrate confidence in the system and rarely seek additional clarification may receive shorter or less intrusive explanations, thereby reducing cognitive load and preserving immersion. In this way, explanation strategies evolve continuously in response to individual user behaviour rather than remaining static throughout the interaction.

An important characteristic of the proposed framework is its recognition that **both over-trust and under-trust can negatively affect user experience and decision quality**. Excessive trust may encourage users to accept inappropriate adaptations without critical evaluation, while insufficient trust may lead users to ignore beneficial recommendations or repeatedly override system decisions. Consequently, the Trust Calibration Loop seeks to maintain an appropriate balance between transparency and user confidence by adapting explanation depth, modality, and frequency according to observed trust levels and system reliability. This adaptive approach supports responsible human–AI collaboration while reducing the risks associated with automation bias and unnecessary user scepticism.

From a long-term perspective, trust calibration should be treated as a **continuous learning process** rather than a single evaluation stage. As users gain experience with adaptive VR systems, their expectations, confidence, and explanatory needs naturally evolve. Accordingly, the framework recommends periodically updating trust models using longitudinal interaction data while ensuring that users retain meaningful control over explanation preferences and adaptive behaviours. Such continuous refinement enables the system to remain responsive to changing user requirements while maintaining transparency and consistency across repeated VR sessions.

Publisher
https://www.ijcsejournal.org/ Type of Article (Review Article)

4.7 Evaluation Layer

The **Evaluation Layer** provides a comprehensive mechanism for assessing whether explainable adaptive Virtual Reality (VR) systems achieve their intended objectives without compromising the quality of the immersive experience. Unlike conventional Explainable Artificial Intelligence (XAI) applications, where evaluation primarily focuses on explanation accuracy and interpretability, adaptive VR requires a multidimensional assessment that simultaneously considers technical performance, user experience, trust, and immersive interaction. Consequently, the proposed framework recommends evaluating explainability using both **AI-centred** and **VR-specific** performance indicators to ensure that transparency enhances rather than diminishes the overall effectiveness of adaptive systems.

From an XAI perspective, evaluation should examine the quality of generated explanations using established criteria such as **explanation fidelity**, **interpretability**, **comprehensibility**, **completeness**, and **usefulness**. Fidelity measures the extent to which explanations accurately represent the reasoning of the underlying AI model, whereas comprehensibility evaluates whether users can correctly understand and interpret the explanation. Additional measures, including explanation satisfaction, perceived transparency, and perceived fairness, provide valuable insights into the quality of human–AI interaction beyond purely technical performance.

Because adaptive decisions directly influence immersive experiences, the framework further incorporates **VR-specific evaluation metrics**. User presence and immersion should be assessed using validated instruments such as the **Igroup Presence Questionnaire (IPQ)**, while user comfort and cybersickness can be evaluated through the **Simulator Sickness Questionnaire (SSQ)**, as demonstrated in previous adaptive VR research (Martin et al., 2020; Sameri et al., 2024). Complementary measures, including task completion time, task accuracy, learning outcomes, cognitive workload (e.g., NASA-TLX), and overall usability, provide additional evidence regarding whether explainable adaptations improve user performance without increasing mental effort or reducing engagement.

A distinguishing feature of the proposed framework is that **trust calibration is treated as a measurable evaluation outcome rather than an assumed consequence of explainability**. Evaluation should therefore examine whether explanations improve appropriately calibrated trust, encourage informed reliance on adaptive recommendations, and reduce both automation bias and unnecessary user scepticism. Behavioural indicators, such as adaptation acceptance rates, explanation requests, manual overrides, and long-term system usage, should be analysed alongside subjective trust questionnaires to provide a comprehensive assessment of user confidence in adaptive AI systems.

The framework also emphasizes the importance of evaluating **ethical and user-centred dimensions** of explainability. Future validation studies should investigate users' perceptions of privacy, informed consent, transparency, fairness, accessibility, and perceived control over adaptive behaviour. These factors are particularly important in immersive environments that continuously collect behavioural and physiological data, where responsible AI principles extend beyond algorithmic performance to encompass user autonomy and data governance.

Rather than relying on a single performance indicator, the proposed framework advocates a **multidimensional evaluation strategy** that integrates technical, behavioural, experiential, and ethical outcomes. An explanation strategy that improves explanation accuracy but significantly reduces presence, increases cognitive workload, or negatively affects user comfort should be regarded as unsuccessful, since transparency should enhance rather than undermine the immersive qualities that define effective VR experiences.

5. Discussion

The proposed seven-layer framework advances current research by integrating Explainable Artificial Intelligence (XAI) principles directly into the architecture of adaptive Virtual Reality (VR) systems rather than treating

Publisher
https://www.ijcsejournal.org/ **Type of Article** (Review Article)

explainability as a post hoc visualization or interface component. Existing research has largely progressed along two parallel trajectories: XAI has concentrated on improving the interpretability of machine learning models, whereas adaptive VR has primarily focused on enhancing personalization, responsiveness, and user engagement (Guidotti et al., 2018; Li et al., 2025). The proposed framework bridges these research streams by embedding explainability throughout the adaptive decision-making lifecycle, from data acquisition and inference to explanation delivery, trust calibration, and multidimensional evaluation. This integration represents the primary theoretical contribution of the framework.

One of the most significant design challenges addressed by the framework is the **balance between transparency and immersion**. Unlike conventional XAI applications, where additional explanations generally improve user understanding with limited negative consequences, explanations in immersive environments may interrupt user attention, reduce presence, or increase cognitive workload. Rather than viewing this trade-off as an unavoidable limitation, the proposed framework explicitly models it through the **Explanation Modality** and **Granularity and Timing Control** layers. By allowing explanation strategies to vary according to adaptation importance, user preferences, and interaction context, the framework supports adaptive transparency that seeks to maximize understanding while preserving immersion (Rahimi et al., 2025).

A second important consideration concerns the **computational demands of real-time explainability**. Adaptive VR systems require rapid decision-making to maintain responsive interaction, whereas advanced explanation techniques such as SHAP may introduce significant computational overhead (Lundberg & Lee, 2017). The proposed framework therefore recommends hybrid inference architectures in which computationally intensive adaptive models are complemented by lightweight interpretable components capable of providing timely explanations. Such architectures, consistent with emerging trends identified by Li et al. (2025), represent a practical compromise between predictive performance and explanation efficiency. Future empirical research should investigate the relationship between explanation fidelity, computational latency, and user-perceived explanation quality under real-time operating conditions.

The framework also highlights the importance of **personalized explainability**. Users differ considerably in their technical expertise, cognitive preferences, and desired level of interaction with intelligent systems. Consequently, a single explanation strategy is unlikely to satisfy all users or all application contexts. The proposed Trust Calibration Loop enables explanation strategies to evolve dynamically according to user behaviour, explanation requests, and observed trust levels. Over time, this adaptive approach may reduce explanation fatigue while improving calibrated trust and long-term user acceptance. Such personalization extends current XAI research beyond static explanation generation toward continuous human–AI collaboration.

Another important contribution of the framework is its integration of **ethical and privacy considerations** into adaptive VR design. Because many adaptive systems rely on behavioural and physiological sensing, users may reasonably expect transparency regarding what information is collected, how it is interpreted, and how it influences adaptive decisions. The framework therefore supports responsible AI principles by promoting informed consent, user awareness, selective disclosure of adaptation mechanisms, and privacy-conscious handling of sensitive biometric data. These considerations extend explainability beyond algorithmic transparency to encompass responsible data governance and user autonomy, both of which are increasingly recognised as essential components of trustworthy AI.

From a practical perspective, the proposed framework provides implementation guidance for researchers and developers across multiple application domains. In educational VR, explanations can improve learner understanding of adaptive instructional strategies and encourage self-regulated learning. In healthcare and rehabilitation, transparent adaptation may strengthen patient confidence in AI-assisted therapeutic interventions while improving clinician oversight. Industrial training environments may benefit from explanations that justify adaptive difficulty adjustments and personalized feedback, thereby enhancing learning effectiveness and user confidence.

Publisher
https://www.ijcsejournal.org/ **Type of Article** *(Review Article)*

Entertainment applications may similarly employ adaptive explanation strategies that preserve immersion while increasing player awareness of AI-driven personalization. Consequently, the framework offers a flexible foundation that can be adapted to diverse immersive applications.

Despite these contributions, several limitations should be acknowledged. The framework is conceptual and has not yet been empirically validated through controlled user studies or real-world implementations. Furthermore, although the framework is designed to remain independent of specific AI algorithms, different machine learning models may require specialized explanation techniques that were not examined in detail. Similarly, application domains such as healthcare, education, and industrial training impose distinct regulatory, ethical, and usability requirements that may influence explanation design. Future research should therefore validate the framework across diverse VR platforms and user populations while investigating the effects of explanation modality, timing, personalization, computational efficiency, and long-term trust calibration on user experience and system performance.

Overall, the proposed seven-layer conceptual framework establishes a comprehensive approach for integrating Explainable Artificial Intelligence into adaptive Virtual Reality systems. Unlike existing approaches that often treat explainability as an auxiliary feature, the framework embeds transparency throughout the adaptive lifecycle, encompassing user sensing, intelligent decision-making, explanation generation, adaptive delivery, trust calibration, and multidimensional evaluation. By combining technical, human-centred, and ethical perspectives, the framework contributes to the development of adaptive VR systems that are not only intelligent and personalized but also transparent, trustworthy, and user-centric. Furthermore, it provides a flexible conceptual foundation for future research and practical implementation across diverse domains, including education, healthcare, rehabilitation, industrial training, and immersive entertainment. Future empirical studies should validate the proposed framework through user-centred experiments and real-world deployments to assess its effectiveness, scalability, and generalizability across different VR applications.

6. Limitations and Future Work

Although the proposed seven-layer framework offers a comprehensive conceptual approach for integrating Explainable Artificial Intelligence (XAI) into adaptive Virtual Reality (VR) systems, several limitations should be acknowledged. First, the framework is conceptual in nature and has not yet been validated through empirical implementation or user-based experimentation. Consequently, the relationships among the proposed layers, as well as the effectiveness of different explanation strategies, remain theoretical and require experimental verification. In particular, the proposed recommendations regarding explanation modality, granularity, and timing should be evaluated through controlled user studies that compare alternative explanation strategies across identical adaptive VR scenarios.

Second, the framework has been synthesized primarily from research on adaptive personalization, cybersickness prediction, exposure therapy, and explainable AI (Martin et al., 2020; Kritikos et al., 2021; Li et al., 2025; Rahimi et al., 2025). Although these domains provide a strong foundation for conceptual development, additional research is needed to determine the framework's applicability to other immersive contexts, including collaborative social VR, industrial training, remote education, cultural heritage, and large-scale metaverse environments. Different application domains may impose distinct requirements regarding explanation timing, user interaction, safety, and privacy, which could necessitate domain-specific adaptations of the framework.

A further limitation concerns the computational requirements associated with real-time explainability. Advanced explanation techniques such as SHAP and counterfactual reasoning may introduce latency that is incompatible with resource-constrained standalone VR headsets. While the framework recommends hybrid architectures that balance

Publisher
https://www.ijcsejournal.org/ Type of Article (Review Article)

predictive performance and interpretability, future research should empirically investigate the trade-offs among explanation fidelity, computational efficiency, energy consumption, and real-time responsiveness across different hardware platforms.

Another important area for future investigation involves **personalized explainability**. The present framework assumes that explanation preferences vary across users and contexts; however, it does not specify adaptive algorithms for learning these preferences over time. Future studies should explore intelligent explanation personalization based on user expertise, cognitive workload, interaction history, trust levels, and accessibility requirements. Such adaptive mechanisms could further improve user satisfaction while reducing explanation fatigue and unnecessary cognitive interruptions.

Future research should also strengthen the empirical evaluation of the framework through longitudinal user studies that examine transparency, trust calibration, immersion, usability, cognitive workload, learning outcomes, and technology acceptance across diverse VR applications. Comparative evaluations against existing adaptive VR architectures would further establish the practical benefits and limitations of embedding explainability throughout the adaptive pipeline.

Finally, this paper represents a conceptual synthesis rather than a formal systematic review. Although the framework was developed by integrating evidence from explainable AI, adaptive VR, and trust calibration research, the literature selection did not follow a predefined systematic review protocol. Future work could therefore complement this conceptual contribution with a **PRISMA-guided systematic literature review** (Page et al., 2021) or a meta-analysis to provide a more comprehensive evidence base and further refine the proposed framework.

Overall, these limitations do not diminish the contribution of the proposed framework; rather, they define a clear research agenda for advancing transparent, trustworthy, and human-centred adaptive Virtual Reality systems. By combining empirical validation, algorithmic innovation, user-centred evaluation, and domain-specific implementation, future studies can extend the framework from a conceptual model into a practical foundation for next-generation explainable immersive environments.

7. Conclusion

Adaptive Virtual Reality (VR) systems are increasingly capable of delivering personalized experiences by continuously interpreting behavioural, physiological, and contextual user data. However, the growing use of intelligent adaptation has introduced significant challenges regarding transparency, trust, and user understanding, as existing Explainable Artificial Intelligence (XAI) approaches were primarily developed for static and non-immersive decision-making environments (Guidotti et al., 2018; Ribeiro et al., 2016; Lundberg & Lee, 2017). This highlights the need for explainability mechanisms specifically designed to address the unique characteristics of adaptive immersive systems.

To address this challenge, this paper proposed a **seven-layer conceptual framework** that embeds explainability throughout the adaptive VR lifecycle, encompassing adaptation triggers, intelligent inference, explanation generation, explanation delivery, granularity and timing control, trust calibration, and multidimensional evaluation. Rather than treating explanations as an interface-level addition, the framework positions explainability as a fundamental architectural principle that supports transparent, trustworthy, and human-centred adaptive VR systems. By integrating technical, behavioural, experiential, and ethical considerations within a unified framework, the proposed model provides a structured foundation for both the design and evaluation of explainable adaptive VR applications.

Publisher
https://www.ijcsejournal.org/ **Type of Article** *(Review Article)*

The proposed framework contributes to the emerging intersection of XAI and immersive computing by extending explainability beyond algorithmic interpretation toward user-centred interaction, adaptive communication, and continuous trust calibration. Its model-agnostic design enables broad applicability across diverse domains, including education, healthcare, rehabilitation, industrial training, and immersive entertainment, while providing practical guidance for researchers and developers seeking to incorporate explainability into next-generation adaptive VR systems.

Although the framework remains conceptual, it establishes a clear research agenda for future work. Empirical validation through user-centred experiments, prototype implementations, and longitudinal evaluations will be essential to examine the effectiveness of different explanation strategies, modality selection, adaptive personalization, and trust calibration mechanisms. Such investigations will help refine the proposed framework and advance the development of adaptive VR systems that balance personalization, transparency, ethical responsibility, and immersive user experience.

Ultimately, this work demonstrates that explainability should not be regarded as an auxiliary feature of adaptive Virtual Reality but as a **core design principle** for intelligent immersive systems. Embedding transparency throughout the adaptive pipeline has the potential to strengthen user trust, improve decision understanding, and support the development of more responsible, effective, and human-centred Virtual Reality experiences.

References

2009

Publisher
https://www.ijcsejournal.org/ **Type of Article** *(Review Article)*

Lim, B. Y., & Dey, A. K. (2009). *Assessing demand for intelligibility in context-aware applications*. In *Proceedings of the 11th International Conference on Ubiquitous Computing* (pp. 195–204). Association for Computing Machinery. https://doi.org/10.1145/1620545.1620576

2016

Ribeiro, M. T., Singh, S., & 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* (pp. 1135–1144). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778

2017

Lundberg, S. M., & Lee, S.-I. (2017). *A unified approach to interpreting model predictions*. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 4765–4774).

2018

Adadi, A., & Berrada, M. (2018). *Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). IEEE Access, 6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Pedreschi, D., & Giannotti, F. (2019). *A survey of methods for explaining black box models. ACM Computing Surveys, 51*(5), Article 93. https://doi.org/10.1145/3236009

2019

Alghofaili, R., Sawahata, Y., Huang, H., Wang, H.-C., Shiratori, T., & Yu, L.-F. (2019). *Lost in style: Gaze-driven adaptive aid for VR navigation*. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–12). Association for Computing Machinery. https://doi.org/10.1145/3290605.3300850

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI—Explainable Artificial Intelligence. *Science Robotics, 4*(37), eaay7120. https://doi.org/10.1126/scirobotics.aay7120

Miller, T. (2019). Explanation in Artificial Intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

2020

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Martin, N., Mathieu, N., Pallamin, N., Ragot, M., & Diverrez, J.-M. (2020). Virtual reality sickness detection: An approach based on physiological signals and machine learning. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)* (pp. 387–399). IEEE. https://doi.org/10.1109/ISMAR50242.2020.00065

2021

Publisher
https://www.ijcsejournal.org/ **Type of Article** *(Review Article)*

Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O., & Weisz, J. D. (2021). Expanding explainability: Towards social transparency in AI systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Article 82). Association for Computing Machinery. https://doi.org/10.1145/3411764.3445188

Kritikos, J., Alevizopoulos, G., & Koutsouris, D. (2021). Personalized virtual reality human-computer interaction for psychiatric and neurological illnesses: A dynamically adaptive virtual reality environment that changes according to real-time feedback from electrophysiological signal responses. *Frontiers in Human Neuroscience, 15*, 596980. https://doi.org/10.3389/fnhum.2021.596980

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

2022

Cancro, G. J., Pan, S., & Foulds, J. (2022). *Tell me something that will help me trust you: A survey of trust calibration in human-agent interaction*. arXiv.

2024

Sameri, M. J., Coenegracht, H., Van Damme, S., De Turck, F., & Torres Vega, M. (2024). Physiology-driven cybersickness detection in virtual reality: A machine learning and explainable AI approach. *Virtual Reality, 28*(4), 174. https://doi.org/10.1007/s10055-024-01067-z

2025

Li, T., Zhu, Y., Liang, H.-N., & Wang, Y. (2025). *From adaptation to intelligence: A systematic review of data, strategies, and impact in personalized VR*. arXiv.

Newen, C., Bodemer, D., Glantz, S., Müller, E., Wischnewski, M., & Schnaubert, L. (2025). *Uncertainty awareness and trust in Explainable AI: On trust calibration using local and global explanations*. arXiv.

Rahimi, F., Sadeghi-Niaraki, A., & Choi, S.-M. (2025). Generative AI meets virtual reality: A comprehensive survey on applications, challenges, and future directions. *IEEE Access, 13*, 94893–94909. https://doi.org/10.1109/ACCESS.2025.3574779