Islamic University Journal of Applied Sciences (IUJAS)
https://journals.iu.edu.sa/jesc

Volume VII, Issue II, December 2025, Pages 120-136

Balancing Personalization and Transparency in User-Centered AI Systems Through Explainable Deep Learning Interfaces

Ahmed Alshehri [1*]

[1] Department of Information Technology, Faculty of Computing and Information, Al-Baha University, Al-Baha, Saudi Arabia. (a.alyehyawi@bu.edu.sa)

*Corresponding author: (Ahmed Alshehri), *Email Address*: a.alyehyawi@bu.edu.sa

Abstract

As AI systems become more advanced and personalized for user experiences in multiple contexts, such as e-learning, finance, and healthcare, the need and necessity for transparency become even greater. While deep learning models can provide the highest quality and recommendations, their black-box nature can inhibit user understanding, trust, and control. In this work, we explore the balance between personalization and transparency in user-centered AI systems by including explainable AI (XAI) techniques in deep learning algorithm-based recommender systems. Our study provides a system with a hybrid architecture that models user behavior embeddings, LSTM/CNN layers, and attention-based mechanisms. Explanations were provided for users through SHAP values, attention-based visual cues, and natural language text that helped users interpret their recommendations in real time. The interface with visual overlays and user interactive panels were designed for the user as a function of cognitive load and types of explanation. The proposed system was tested through two phases of user studies, both with quantitative performance metrics and qualitative data. The results indicated better recommendation accuracy, trust, perceived fairness, and user satisfaction when users received explanations. This work indicates how we can build ethical and usable AI systems. We show that by employing explainable interfaces we can not only enhance the effectiveness of personalized technology, but also increase human-level acceptability.

Keywords: UX, XAI; Personalization; User-Centered AI; Deep Learning; Transparency.

https://doi.org/10.63070/jesc.2025.026
Received 31 August 2025; Revised 01 October 2025; Accepted 08 October 2025.
Available online 12 October 2025.

1. Introduction

The mass proliferation of artificial intelligence (AI) into consumer facing applications has radically transformed the relationship users have with digital systems. Personalization is virtually ubiquitous in today's applications, from personalized learning environments and streaming services to healthcare and financial services; AI systems increasingly personalize both the content and decision-making to suit individual user preferences and behaviours [1]. These capabilities have significantly enhanced user engagement and satisfaction, and are increasingly powered by deep learning models. Despite their overall benefits, however, there is an important limiting factor - these systems often provide limited transparency, leaving users unclear as to how decisions are arrived at, or why specific recommendations are offered [2]. This opacity undermines user trust, can reduce perceived control, and obstructs the development of an accurate mental model the user may need for confident interaction. This research investigates the core challenge of achieving a balancing act of personalization and transparency in user-centered AI systems. Deep learning models can provide highly accurate, contextual, and reliable recommendations, even if those recommendations are not always reasonable to the user, but they often function in a black box manner that is opaque to the user, rendering the system unexplainable, and difficult for the user to interpret or question [3]. In contrast, interpretable models will generally lend themselves to more transparency, but they will trade-off performance and can often provide some transparency, but at a cost to the performance. This design problem is important to consider when developing AI applications, especially in domains in which user trust, user autonomy, and ethics are critical.

To anchor this research in recognized user experience (UX) and human-computer interaction (HCI) theory, the research introduced several established foundations. Don Norman's Seven Stages of Action, prompts us that the human-computer interaction is the distance between the system's output and user intent [4]. For users to feel in control and make decisions, they need to get a sense of the AI system's behavior including perceiving, interpreting and evaluating. It is not a surprise that when explanations are not present or are unintelligible, this alignment is lost. Lee and See Trust in Automation framework, introduces that "trust in machine learning process is dynamic and need to be calibrated to display effective trust and reduce excess trust" [5]. The extent to which a user is blind to automation depends on the transparency of the system, which aids the user in developing the right amount of trust where they are neither blindly reliant on nor unjustifiably sceptical of the automated process. Additionally, principles from the Cognitive Load Theory to avoid cognitive overload from complexity and over-Information [6]. To be sufficiently informative and helpful enough to meet the information needs of

users, explanations must allocate concise cognitive load that is just enough to be helpful, but not so much as to be unmanageable.

To address these individual concept challenges, this research introduced a hybrid deep learning framework capable of reporting accurate personalized recommendations but simultaneously including aspects of XAI for user understanding and promoting trust. The proposed system utilizes embedding and sequence-based models to recognize user behavioral patterns and user preferences. By including SHAP (SHapley Additive exPlanations) [7] values to show feature contributions, and attention-based approaches to surface important behavioral indicators for recommendations, the design addresses the need for transparency. The objective is to provide a way to explain recommendations from the recommendation system with visual (e.g., charts/heatmaps) and textual (e.g., natural language summaries) explanations for users with varying levels of technical knowledge.

The framework is evaluated by using a comprehensive user study that measures both objective performance measures (e.g., the accuracy of recommendations) and subjective measures (e.g., trust, satisfaction, and perceived transparency) including how users engage with explainable recommendation interfaces and how explanations influence their perceptions of the system. By studying these actions and situations within a real-world application space, the research will help examine the nuanced relationship of personalization and interpretability.

Altogether, the work contributes to designing ethical, user-aligned AI systems by showcasing that explainable deep learning-based interfaces improve not just the performance of the system but also the trust, satisfaction, and engagement of users. This contributes to the goal of informing further development of transparent human-centered AI systems which promote informed decision-making and responsible automation.

2. Background

The growth of interactive AI systems in several sectors including e-learning, digital marketing, healthcare, and entertainment has vastly improved user experience via personalized content recommendation, contextual relevance and task performance [8]. These systems use user information to better recommend, predict and decide within a normal user's preferences for a more relevant and engaging experience.

Deep learning has fueled personalization as it can best represent complex user-item interaction, temporal behaviour and high-dimensional feature space [9]. The growth of interactive AI systems in several sectors including e-learning, digital marketing, healthcare, and entertainment has vastly improved user experience via personalized content recommendation, contextual relevance and task performance [8]. These systems use user information to better recommend, predict and decide within

a normal user's preferences for a more relevant and engaging experience. Deep learning has fueled personalization as it can best represent complex user-item interaction, temporal behaviour and high-dimensional feature space [9]. Nonetheless, even when successful, personalization systems powered by deep learning have a common challenge: complexity has rendered them opaque. Users generally are not able to understand the logic leading to an AI-enabled decision, especially in high-stakes or sensitive domains. Impaired trust and confidence is not the only concern; it may also limit users' ability to make rational, informed decisions or challenge what the system may output [10]. To deal with this issue, the new field of explainable AI (XAI) [11], has started to present methods to make AI decisions interpretable and accessible for end users.

Explainability is a design issue, in human-computer interaction (HCI) terms, not just a technical issue. It is a design issue in which user psychology and cognition are critically important [12]. Benefits of a user-centered design approach, and appropriate UX and HCI normative theories, systems must provide the user with outputs that they can see and evaluate with respect to their goals and expectations [13]. If a user cannot understand why a system made a recommendation, then they cannot build a correct mental model, which is affecting how usable the system is. Similarly, the trust in AI systems which can be modeled with frameworks, such as Lee and See's Trust in Automation, - requires that the AI system provide justification which manages requirements for trust levels (e.g., calibration of trust) [14]. Cognitive load theory advises caution against providing overly elaborate details or technical explanations that exceed a user's cognitive processing capacity, suggesting a need for lightweight, contextualized, and actionable explanations. As such, the implementation of explainability with AI systems (e.g., personalization) must be a transdisciplinary problem involving machine learning, user-centered design, and cognitive science.

3. Related Work

The intersection of personalization and transparency in user-centered AI systems is still an active area of research, especially in developing explainable deep-learning interfaces that inform explanation, generate trust and understanding. Some recent literature has emphasized the important design task of incorporating personalization into user-centered AI-based development while still being able to produce clear and interpretable explanations. A study by [15], highlights the importance of user-centered explainability in the context of energy demand forecasting in smart homes, emphasizing that personalized explanations for the end-user enhance usability and trust in the system. Authors [16], proposed a human-centric approach to personalization in AI for education, suggesting a multimodal modular architecture and an interpretable modular system that enables the model to choose explanations according to user's context, thus making a trade-off between personalization and

transparency. This approach demonstrates that adaptive explanations can be generated for many different groups of users despite the complexity of each group, thus enhancing understanding and engagement. In the healthcare, the need for transparent AI models is relevant because many users prefer an interoperable rationale AI model. A recent review on explainable AI in healthcare notes that if an AI model is transparent for example, it allows the user to build trust and better decision-making especially if it can explain how it makes decisions [17]. Complementing this, AI's role in hospitals and clinics demonstrates how explainability facilitates an improvement in traditional systems, such as finding meaning within high-dimensional data, that leads to clinical decision-making [18]. Educational settings also provide motivation for explainable AI. The review on AI in education notes how explainability can help adaptive learning systems improve decision support. Educators and students can improve their understanding of recommendations, feedback, and learning pathways stemming from the use of AI systems [19]. Other complexities in this literature address AI systems' adaptability, with a multi-layered framework for research advocating safe and personalized explainability, with recommendations for future systems to modify the explanations given based upon individual user preferences and knowledge, making systems customizable for the end user through trustworthiness and usability [20]. Much of the literature concerning an explanation's approximation of black-box models lays claim to unintentional consequences on developing explainability toward traditional AI system users. Nevertheless, other recent reviews have noted the intent of explainability methods to generate explainable outcomes that are readable, usable, and provide users understanding and trust. These user-interface design methods cross domains, often in regard to trust and understanding, as associated with acceptance processes and improved understanding of AI reasoned thinking within domains including healthcare and education [21, 22]. There is clear divergence in the design approach towards intuitive, user-centered interfaces, intended to design transparent AI applications that align with human-centered design principles toward usability and trustworthiness as a primary consideration in designing explainability [23]. Additionally, the development of explainability and comprehensibility is proven to be essential in developing user trust in deep learning systems. There are a variety of discoveries that explain the production of a decision, based on statistical pattern recognition methods, that can help make the output of AI more comprehensible to foster acceptance [24]. The concept of human-centered AI (HCAI) directly endorses systems that provide clear insight into their reasoning process so that users can understand and assess AI's decisions [25]. Lastly, design research on the usability and trustworthiness of AI-based interfaces is on the rise, with studies suggesting that explainability increases user confidence and acceptance of the system. Studies from multiple disciplines suggest that designing AI interfaces with explainability not only creates transparency, but also aligns with user expectations for having a benevolent and trustworthy

personalized AI [26]. Specifically, the body of literature examined shows that creating a user-centered AI system requires balancing personalization and transparency in the adequacy of designing user centered AI systems requires creating adaptable but interpretable explanations, which balance different user needs while ensuring the user understands AI's reasoning process. Ultimately, the balance will be necessary for trust, usability, and humans and AI system collaboration across various application areas.

4. System Architecture and XAI-Driven Interface Design

The proposed system consists of human-centered, explainable deep learning architecture that captures accuracy of user-centered personalized recommendations with transparent, interpretable justification to users. The architecture consists of three parts as given in Fig (1): (1) hybrid deep learning model for user behavior learning, (2) an explainability layer for generating real-time human-understandable explanations, and (3) a user interface that presents explanation to the user in visual and textual form.

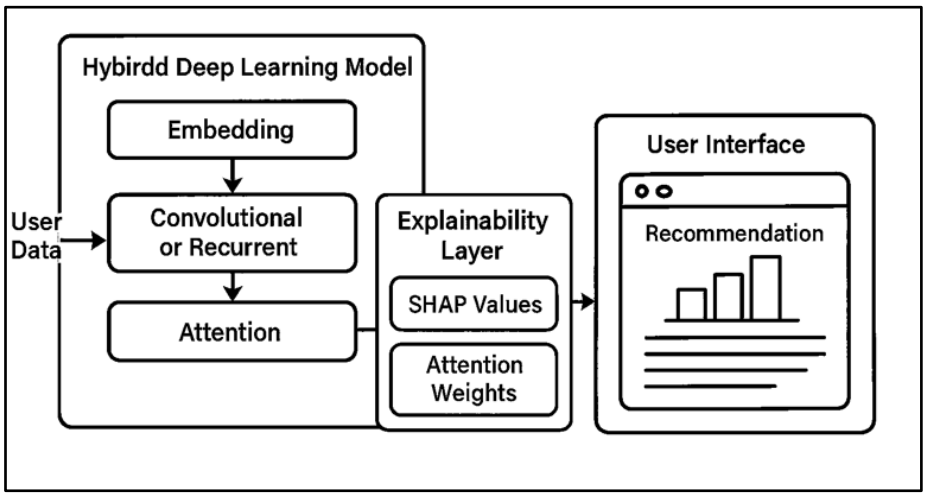

Fig 1: Proposed Approach

The architecture uses user interaction data as input including the user ID, item ID, user engagement history, and item content metadata. These inputs are embedded into dense vectors, and then passed through CNN and LSTM layers to learn spatial and temporal patterns in user behavior. An attention mechanism then aggregates the representations by assigning weights to the most important user interactions, which produces the final recommendation output according to either ranked items or predicted preferences. The explainability layer will run in tandem and will provide SHAP-based feature attributions and attention maps that highlight important pieces of evidence, when possible. The recommendation and explainability outputs will then be combined, converting each into visual and

textual explanations, through the explanation engine, in a way that allows the user to visualize both the prediction and the reasoning behind the prediction through the interface.

4.1 Hybrid Deep Learning Model

The system employs a hybrid deep learning model—including representations of both static user features and dynamic behavior patterns so it can move toward high personalization performance. The model architecture contains embedding layers, convolutional, recurrent layers, and an attention layer. The embedding layers convert categorical input to dense vector representations for instances that have user IDs, item IDs, and/or user-enhanced content metadata elements. The goal is to learn the latent associations between user, and item. From these embeddings each of which are of a specified embedding length, the layers are either sent into Long Short-Term Memory (LSTM) units or 1D Convolutional Neural Networks (CNN) units based on the nature of input. LSTM model it is important for capturing temporal sequences history of temporal/user engagement such as the order of accessed learning resources or user engagement (time-stamped history). The CNN model is best suited to recognize local patterns in the user's interaction (fixed length sequences) or the nature of features in the content descriptors. In either setup the network is trained to learn associations between a set of behavioral patterns and their likely user preference behaviors. An attention mechanism is included in the last stage of the network to improve interpretability as well as performance. The attention module will assign a relevance score for each element in the input (e.g., previous user interaction) relative to the final output prediction. This affords the model an opportunity to emphasize the most contextually relevant elements and provides a natural incorporation point for explanations, as attention weights can simply be visualized or used to rank inputs by contribution Algorithm 1 provides the technical details of the proposed approach.

The combination of CNN, LSTM, SHAP, and attention mechanisms improves accuracy and interpretability. CNN layers extract local spatial features from user–item interactions, and LSTM layers model temporal dependencies in the user behavior. The attention mechanism gives relevance weights to user interactions in order to highlight the most important user–item interactions, providing an interpretive bridge between CNN and LSTM outputs. Then, we use SHAP for post-hoc feature attribution, providing clear transparency along multiple levels of the recommendation process.

4.2 Explainability Layer

Although the hybrid deep learning model is strong in terms of prediction strength, it has interpretability through the explainability layer. The explainability layer gives some information to the user about the

recommendation that is made by combining model-specific and model-agnostic methods in near real time. The first part of the explanation is the use of SHAP (SHapley Additive exPlanations) values, which is a model-agnostic way to estimate the contribution of input features to the final prediction. SHAP has global and local interpretability: it can tell you which features are generally important across users and it is also able to explain the recommendations for a particular user by showing the importance of the input feature. The second part of the explanation is the use of the model's attention mechanism. During training, the model will learn what the attention weights mean and this allows the model to make recommendations. The attention weights show what parts of the interaction history for a user mattered for the recommendation, as a measure of importance. Careful extraction of this information will provide additional supporting visual cues in the user interface. For example, the prompts can be presented by highlighting elements, intensities of bars or colors, along timelines, and so forth. The next stage is the textual explanation generator that parses SHAP and attention layer technical outputs into natural language statements. These statements are produced by pre-existed templates, which are filled out dynamically, depending on the actual inputs and the model outputs. For example: "This module is recommended because you recently interacted with topics similar to data privacy and cyber security". The aim is to produce concise, human-readable explanations that follow user's mental models and cognitive capabilities.

By combining SHAP with attention mechanisms, we create a multi-level framework for explainability. The attention layer adds an inherent level of interpretability by presenting salient interactions. SHAP features, on the other hand, provide model-agnostic feature attributions, in terms of significance, after a prediction has been made. Both techniques work in conjunction to create more transparency, by providing attention-based descriptions of influence with SHAP-based feature attributions, in a visual and text-based output, to increase understanding for the user.

Figure 2 presents the proposed Algorithm 1 that describes the hybrid recommendation process. Features for users and items are embedded and passed through CNN and LSTM layers to extract and learn spatial and temporal features. Attention assigns weights to important interactions and outputs recommendations at the recommendation stage. In the explainability stage, we calculate SHAP values for feature attribution, and examine important interactions using the attention weights to prompt potential behavior evidence. The explanation engine combines the attribution and important nature of interactions to provide textual and visual explanations to users.

Algorithm 1 Proposed Algorithm

Notation:
Embed$(\cdot)$: Embedding layer
θ: Model parameters
$\sigma(\cdot)$: Sigmoid activation
procedure HybridModel$(\mathbf{u}, \mathbf{X}, \mathbf{i})$
  $\mathbf{e}_u \leftarrow \text{Embed}_{\text{user}}(u)$ ▷ User embedding
  $\mathbf{E}_{\text{seq}} \leftarrow \text{Embed}_{\text{seq}}(\mathbf{X})$ ▷ Sequence embedding
  if using LSTM **then**
    $\mathbf{h}_t \leftarrow \text{LSTM}(\mathbf{E}_{\text{seq}}; \theta_{\text{LSTM}})$
  else
    $\mathbf{h}_t \leftarrow \text{CNN}(\mathbf{E}_{\text{seq}}; \theta_{\text{CNN}})$
  end if
  $\alpha_t \leftarrow \text{softmax}(\mathbf{W}_a[\mathbf{h}_t; \mathbf{e}_u])$ ▷ Attention weights
  $\mathbf{c} \leftarrow \sum_{t=1}^{T} \alpha_t \mathbf{h}_t$ ▷ Context vector
  $s(u, i) \leftarrow \sigma(\mathbf{W}_s[\mathbf{c}; \mathbf{e}_u; \mathbf{e}_i] + b_s)$ ▷ Scoring
  $\mathbf{R} \leftarrow \text{topk}_i\, s(u, i)$ ▷ Top-k recommendations
  return $\mathbf{R}, \alpha_t$
end procedure
procedure Explainability$(\mathbf{R}, \mathbf{u}, \mathbf{X})$
  $\phi_j \leftarrow \text{SHAP}(s, \mathbf{u}, \mathbf{i}, \mathbf{X})$ ▷ Feature contributions
  $\mathbf{E}_{\text{text}} \leftarrow \text{Template}(\phi_j, \alpha_t, \mathbf{X})$ ▷ Natural language
  $\mathbf{E}_{\text{visual}} \leftarrow \text{Render}(\alpha_t, \phi_j)$ ▷ Heatmaps/bars
  return $\mathbf{E}_{\text{text}}, \mathbf{E}_{\text{visual}}$
end procedure
procedure Interface$(\mathbf{u}, q)$
  $\mathbf{X} \leftarrow \text{GetInteractions}(u)$
  $\mathbf{R}, \alpha_t \leftarrow \text{HybridModel}(\mathbf{u}, \mathbf{X})$
  $\mathbf{E} \leftarrow \text{Explainability}(\mathbf{R}, \mathbf{u}, \mathbf{X})$
  $\text{Display}(\mathbf{R}, \mathbf{E}_{\text{visual}}, \mathbf{E}_{\text{text}})$
end procedure

Fig.2. The proposed Algorithm 1

4.3 User Interface Prototype

The explainable recommendation system is made available to users in a responsive and intuitive, web-based interface that was designed with well-established human-computer interaction (HCI) principles that prioritize usability, cognitive load, and limited interference with users' regular activities. The front-end incorporates many user-facing components that support interpretability and user control. These include visual overlays (e.g., progress bars, heatmaps, or badges) that elaborate on specific user behaviours or item attributes impacting the recommendations. Such visual indicators enable users to quickly understand how or why a suggestion is being offered without needing a technical background. Additionally, the inclusion of interactivity that triggers tooltips upon a hover indicates something more than a visual representation. The tooltips offer a simple recommendation explanation that is concise and based upon SHAP. Users that would prefer a more elaborate explanation among the user-facing components can access dedicated explanation panels. The explanation panels allow for longer and formatted textual descriptions that articulate the most influential factors that impacted the decision and choices, representing transparency and user confidence. Modern web frameworks such as React.js or

Vue.js are used to develop the frontend component in order to provide modularity, ability to create responsive interfaces, and ease of integration with backend services. The backend architectural includes two connected components, the model inference engine and explanation engine. The model inference engine will load and run the trained hybrid deep learning model to produce real-time, personalized recommendations. The explanation engine will compute SHAP values and an attention map (when relevant to the recommendation) to provide the rationale for the recommendation. These outputs will be formatted dynamically for the visual and textual components of the frontend interface.

In order to maintain the responsiveness of the system, pre-cached explanations are kept in memory for frequently requested queries (in this case, the most common interactions), while all other interactions involve on-demand computation. In other words, a hybrid caching approach will provide flexibility to make trade-offs between performance and responsiveness. The backend capability is designed in a Python-based microservices architecture (Flask, FastAPI) and will be created with common and popular ML libraries (TensorFlow, PyTorch) when serving models, SHAP, and attention. Altogether the frontend and backend capabilities result in a solid architecture that provides both high-quality personalizations and meaningful transparency that is aligned with users.

5. User Study Design

To assess the effectiveness, transparency, and alignment with user views concerning the proposed explainable recommendation system, a two-phase user study was conducted with a combination of qualitative and quantitative components. The study was conducted with the university context of an e-learning platform for learners as they receive recommendations for learning resources, including course modules, video lectures, and additional readings. Using an educational context for the study was important since educational environments require a high level of user trust, perceived equity, and autonomy, creating an environment that could help to evaluate how explainable interfaces impact users' perception of AI positive recommendations.

For the quantitative phase, the participants who were recruited represented a pool of approximately 100 university students living between 18 to 30 years of age, who were currently or recently actively using online learning platforms. The goal of this was to assess the system's performance in real-world use cases, and whether there was an observable effect on user trust, satisfaction, and perceived quality of the system, when using explainable components. It was most important that the participants interacted with a real world interface. The participants were randomly assigned to either of two conditions in an A/B testing context. The control group engaged with a non-explained recommendation

interface, while the experimental group engaged with the explained interface, which included SHAP values, attention based highlights, and text based justifications.

A number of metrics were employed to assess outcomes for both groups. Recommendation effectiveness was evaluated using standard performance metrics: precision, recall, and Normalized Discounted Cumulative Gain (NDCG). User experience was gathered using validated measures, including the Trust in Automation Scale which considers dimensions of trust in systems (i.e., reliability, predictability, and trust in the system), System Usability Scale (SUS), which measures overall usability, and additional custom items to assess subjects' views on transparency, fairness, and understandability.

In addition to the quantitative outcomes, a qualitative evaluation of user cognition and emotions were recorded for a smaller cohort of 15-20 students from a variety of disciplines (including computer science, engineering, and arts) to allow for diverse user feedback regarding the recommendation system which ranged from highly technical users to less knowledgeable users of AI systems. Each participant performed a series of recommendation tasks using the explainable interface while thinking aloud, which was recorded and transcribed for thematic analysis.Upon task completion, participants were engaged semi-structured interviews, to investigate particular aspects of interest relating to system transparency. These aspects included clarity and utility of the explanations, the cognitive effort associated with interpreting and acting upon the feedback from the system, and relevant emotions, including trust, frustration and satisfaction depending on each of a participant's experience. Questions also included whether the explanations created a sense of being in control of their recommendations and whether they would prefer a similar system as part of their daily academic workflows. This combined methodological approach ensured a study which assessed both system effectiveness and user perception, allowing the researchers to develop an overall understanding of how explainability affects user experience, trust and decision-making in a personalized learning experience devoted towards XAI interfaces.

6. Results

The outcomes from the two-phase user study provide strong evidence that using explainable AI techniques within the deep learning recommendation process significantly improves system performance and user experience. The Quantitative results showed that the hybrid deep learning model we proposed (utilizing embedding, LSTM/CNN, and attention capabilities) outperformed the baseline non-explainable recommendation system using standard accuracy metrics. The hybrid models were

found to have higher precision, recall, and Normalized Discounted Cumulative Gain (NDCG) performance levels, indicating it more accurately and contextually understood user preferences. Additional, there was also a statistically significant increase in user trust and satisfaction for people who interacted with the explainable system. The Trust in Automation Scale and the System Usability Scale (SUS) had higher mean scores in the explainable condition while indicating greater confidence and usability in their rating. A summary of the quantitative evaluation results is presented in the Table 1.

Table 1: Quantitative Results Comparison Between Baseline and Explainable Systems

Metric	Baseline System	Explainable System	p-value
Precision	0.81	0.89	< 0.01
Recall	0.78	0.86	< 0.01
NDCG	0.74	0.82	< 0.01
Trust Score	3.10	4.30	< 0.01
SUS Score	68.00	84.00	< 0.01

These results support that the explainable version of the system not only provides better recommendation performance but also significantly enhances the user experience aspects of trust, usability, and perceived transparency. The qualitative results (Table 2) also provided an in-depth exploration of users' strategies relevant to cognitive and emotional engagement. Users reported a clearer representation and correspondence between their mental models of the system behaviors. People reported that the explanations contributed to a better awareness of perceived reasons about the recommendations made by the system. This alignment appeared to support a sense of control and autonomy over users' interactions with the system. Statistical comparisons were performed using independent-sample t-tests with Bonferroni correction ($\alpha = 0.01$). Effect sizes (Cohen’s d) indicated large effects across all metrics, confirming the practical significance of the improvements.

Text based explanations were particularly advantageous for users from non-technical backgrounds. These users found the natural language based justifications to be very accessible and informative. The reasons in natural language allowed users to build context around the recommendations without requiring prior cognitive knowledge or grounding in AI or algorithms. In contrast, users with higher technical proficiency valued the visual explanation elements (e.g., the SHAP-based bar charts and attention highlights) for providing quick, clear, and intuitive insights into what mechanisms contributed to recommendations. In general, users categorized the system as more trustful, transparent, and usable, and preferred interfaces with explainable information. All results support the hypothesis.

Adding explainability to deep learning based recommender systems can increase both the performance of those systems and the ethical use of AI in real-world user-centric applications.

Table 2: Summary of Qualitative Findings from User Study

Theme	Baseline System	Explainable System	User Observation
Mental Model Alignment	Limited	Improved	Users reported better alignment with system behavior
Perceived Control & Autonomy	Low	High	Explanations supported a sense of control
Accessibility of Text Explanations	Mixed Feedback	Very Helpful	Non-technical users found natural language justifications useful
Use of Visual Explanations	Underutilized	Highly Appreciated	Technical users preferred SHAP charts and attention overlays
Overall User Preference	Moderate	Strongly Preferred	Majority preferred the interface with explainable information

7. Discussion

The current results provide meaningful insight into the impact of providing XAI in improving the usability, trustworthiness, and overall effectiveness of deep learning-based recommendation systems. By providing a SHAP-based feature attribution, attention-based visualizations, and textual explanations, it is evident that the supplementary interpretability tools positively impacted usability and endorsed the use of the recommendations, while also improving the overall systems' performance.

One contribution this work, is that providing explanations to recommendations improved the users' cognitive clarity and emotional trust. Individuals in the explainable condition showed substantially improved levels of mental model alignment throughout the task, indicating a better comprehension of the workings of the system and rationale behind the recommendations. It was further evident that interpretability was improved usability and trust in the accompanying system - two aspects critical for system adoption and continued usage of AI technologies. Users indicated a higher likelihood to trust and use the recommendation from the system, expressed a greater sense of control, and classified the interface as fairer reagrding decision making autonomy.

These results are unsurprising, given existing foundations of research in UX and HCI theory. The contributions to our understanding of UX map to Don Norman's Seven Stages of Action, which

suggests that users must be able to perceive, interpret, and evaluate system feedback before they can establish accurate mental models that allow them to successfully achieve their goals. In our study, both types of explanations (visual and textual) provided both perceptual and interpretive bridges between the underlying logic of the system (that a wrong action occurred), and the user expectations. This is a good fit with Norman's Seven Stages of Action.

Our study is also consistent with Lee and See's model of trust in automation. The authors emphasize the importance of calibrated trust - neither over-trust (an unwarranted reliance on the automation), nor under-trust (multiple reasons for suspending trust). By allowing users to read (explanations), understand, and relate (to their own experience) the situation (the explanations must matter to them), we enabled trust to grow out of the user's interaction with the system, rather than out of blind trust in the technology.

The insights also demonstrate important design implications for the future design-explainable AI systems. One of the major takeaways is the need for personalizing the complexity and modality of explanations based on user type. While non-technical users benefitted most from short explanations in natural language, technical users have the tendency to focus mainly on visual cues or required more detailed information in order to gain deeper insight into the reasoning the system's algorithm made. This implies that explanation that is appropriate for one user type is not likely appropriate for the next type. Instead, it would be more effective for a future XAI system to implement adaptive explanation frameworks that would apply explanation content, format, and depth to an individual user's preferred methods of explanation, overall cognitive styles, and domain experience.

Although the study has given very positive results there are some limitations. First, the evaluation was conducted only with university students, and thus, the participants were relatively uniform sampling of a digitally savvy population. While this is appropriate for the educational application domain, it does constrain the generalizability of the findings to other populations, such as older adults, professionals, or even users with lower digital literacy. Future work should draw upon a more diverse group of participants to increase external validity. Second, the application domain was limited to e-learning and while this domain is generally amenable to transparency and control by the user, it may not reflect user behavior in other domains such as e-commerce, healthcare, or social media where the stakes around recommendation and user expectations may be quite different.

From this it is clear that there are many opportunities for future work. One key opportunity is developing adaptive explanation systems that learn from user interactions and adjust the type and

complexity of explanation in real time. Such systems could utilize reinforcement learning or user modeling approaches to culminate the most effective form of explanation based on user needs. Another opportunity is implementing longitudinal studies to find out how sustained explanations affect levels of trust, understand and satisfaction over time. While this study has indicated strong evidence of short-term effects, we cannot ascertain the effect of long term use and adoption of recommender systems that provide explanations. This study has shown that incorporating XAI dimensions in user-centred recommendation systems, enhances technical performance and also embodies core UX and HCI principles, that also contribute to ethical AI. XAI interfaces enhance transparency, provide cognitive support and achieve levels of emotional trust which represent pathways toward usable, ethical, human-aligned AI systems.

8. Conclusion

With the design and evaluation approach, this study provides an integrated framework to tackle the important challenge posed by AI recommender systems: the balance between personalized and transparent recommendations. The study showed that a hybrid deep learning recommender system with built-in explainability (using SHAP-based attributions, attention maps, and natural language explanations) could produce high predictive performance, while providing understandable and trustworthy feedback to users. Using aspects of UX and HCI design principles such as Norman's Action Model and Lee and See's Trust Calibration Framework, the deep-learning recommender produced improvements in cognitive clarity, emotional trust, and reliably confident decision-making, across the diverse university user group in the e-learning context. The quantitative results showed very meaningful improvements in recommendation performance and user satisfaction overall. The qualitative results highlighted the diversity of cognitive needs across the technical and non-technical user group. Visual explanations provided a timely understanding of the recommendations for the experiential users, while textual explanations aided less technical users in building the mental model of the system. This research lays the groundwork for the development of personalized feedback systems that go beyond recommendation quality, and recognize application of user-centered designs recommending flexible and personalized explainability systems, not only in AI applications but where technology, learning, user experience, and user preference convene. The research notes that there are limitations to this research, particularly with respect to the sample population (university students), and contextual limitations (e-learning). Future research would then need to examine the effectiveness of similar explainable systems with different populations, and in increased contexts of examination in much higher stakes applied ethics, such as healthcare and finance. Furthermore, in the area of adaptive

explanation systems, generation of explanations, and longitudinal studies of trust and adoption over time remain ripe for future work. This research enables further development of human aligned AI thinking that explainability is not only a computational additive feature, but rather a central design feature for personalization to be considered ethical, usable, and effective for users becoming empowered with interpretable insight, promoting agency, accountability, and trust, to support and engage in responsible use of AI in pervasive and unchecked ways in everyday life.

References

[1] K. S. Kaswan, J. S. Dhatterwal, and R. P. Ojha, "AI in personalized learning," in *Advances in Technological Innovations in Higher Education*, CRC Press, 2024, pp. 103–117.

[2] S. Patel, R. Patel, R. Sharma, and D. Patel, "Enhancing user engagement through AI-powered predictive content recommendations using collaborative filtering and deep learning algorithms," *Int. J. AI ML Innovations*, vol. 12, no. 3, 2023.

[3] A. Da'u and N. Salim, "Recommendation system based on deep learning methods: a systematic review and new directions," *Artif. Intell. Rev.*, vol. 53, no. 4, pp. 2709–2748, 2020.

[4] D. A. Norman, "Design principles for human-computer interfaces," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 1983, pp. 1–10.

[5] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors*, vol. 46, no. 1, pp. 50–80, 2004.

[6] O. Chen, F. Paas, and J. Sweller, "A cognitive load theory approach to defining and measuring task complexity through element interactivity," *Educ. Psychol. Rev.*, vol. 35, no. 2, p. 63, 2023.

[7] Y. Nohara, K. Matsumoto, H. Soejima, and N. Nakashima, "Explanation of machine learning models using shapley additive explanation and application for real data in hospital," *Comput. Methods Programs Biomed.*, vol. 214, p. 106584, 2022.

[8] D. Gm, R. H. Goudar, A. A. Kulkarni, V. N. Rathod, and G. S. Hukkeri, "A digital recommendation system for personalized learning to enhance online education: A review," *IEEE Access*, vol. 12, pp. 34019–34041, 2024.

[9] G. Gupta and R. Katarya, "Research on understanding the effect of deep learning on user preferences," *Arab. J. Sci. Eng.*, vol. 46, no. 4, pp. 3247–3286, 2021.

[10] T. Miller, I. Durlik, A. Łobodzińska, L. Dorobczyński, and R. Jasionowski, "AI in context: harnessing domain knowledge for smarter machine learning," *Appl. Sci.*, vol. 14, no. 24, p. 11612, 2024.

[11] R. Dwivedi *et al.*, "Explainable AI (XAI): Core ideas, techniques, and solutions," *ACM Comput. Surv.*, vol. 55, no. 9, pp. 1–33, 2023.

[12] Y. Rong *et al.*, "Towards human-centered explainable AI: A survey of user studies for model explanations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 4, pp. 2104–2122, 2023.

[13] L. Van Velsen, G. Ludden, and C. Grünloh, "The limitations of user- and human-centered design in an eHealth context and how to move beyond them," *J. Med. Internet Res.*, vol. 24, no. 10, p. e37341, 2022.

[14] M. Mylrea and N. Robinson, "AI trust framework and maturity model: Improving security, ethics and trust in AI," *Cybersecurity Innov. Technol. J.*, vol. 1, no. 1, pp. 1–15, 2023.

[15] M. Shajalal, A. Boden, and G. Stevens, "Towards user-centered explainable energy demand forecasting systems," in *Proc. 13th ACM Int. Conf. Future Energy Syst.*, 2022, pp. 446–447.

[16] V. Swamy, "A human-centric approach to explainable AI for personalized education," *arXiv preprint*, arXiv:2505.22541, 2025.

[17] S. Bharati, M. R. H. Mondal, and P. Podder, "A review on explainable artificial intelligence for healthcare: Why, how, and when?," *IEEE Trans. Artif. Intell.*, vol. 5, no. 4, pp. 1429–1442, 2023.

[18] S. Maleki Varnosfaderani and M. Forouzanfar, "The role of AI in hospitals and clinics: transforming healthcare in the 21st century," *Bioengineering*, vol. 11, no. 4, p. 337, 2024.

[19] H. Khosravi *et al.*, "Explainable artificial intelligence in education," *Comput. Educ.: Artif. Intell.*, vol. 3, p. 100074, 2022.

[20] R. Oruche, R. Akula, S. K. Goruganthu, and P. Calyam, "Holistic multi-layered system design for human-centered dialog systems," in *Proc. IEEE 4th Int. Conf. Human-Machine Syst. (ICHMS)*, 2024, pp. 1–8.

[21] V. Hassija *et al.*, "Interpreting black-box models: a review on explainable artificial intelligence," *Cogn. Comput.*, vol. 16, no. 1, pp. 45–74, 2024.

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

[23] P. Nama, "AI-powered mobile applications: Revolutionizing user interaction through intelligent features and context-aware services," *J. Emerg. Technol. Innov. Res.*, vol. 10, no. 1, p. g611-g620, 2023.

[24] T. Wischmeyer, "Artificial intelligence and transparency: opening the black box," in *Regulating Artificial Intelligence*, Cham: Springer Int. Publ., 2019, pp. 75–101.

[25] Interaction Design Foundation (IxDF), "What is Human-Centered AI (HCAI)?," *Interaction Design Foundation*, Aug. 5, 2025. [Online]. Available: https://www.interaction-design.org/literature/topics/human-centered-ai

[26] A. Lombardi, S. Marzo, T. Di Noia, E. Di Sciascio, and C. Ardito, "Exploring the usability and trustworthiness of AI-driven user interfaces for neurological diagnosis," in *Adjunct Proc. 32nd ACM Conf. User Modeling, Adaptation and Personalization*, 2024, pp. 627–634.