XAI to Increase the Effectiveness of an Intelligent Pedagogical Agent

John Wesley Hostetter
jwhostet@ncsu.edu
North Carolina State University
Raleigh, North Carolina, USA

Cristina Conati
conati@cs.ubc.ca
University of British Columbia
Vancouver, BC Canada

Xi Yang
yxi2@ncsu.edu
North Carolina State University
Raleigh, North Carolina, USA

Mark Abdelshiheed
mnabdels@ncsu.edu
North Carolina State University
Raleigh, North Carolina, USA

Tiffany Barnes
tmbarnes@ncsu.edu
North Carolina State University
Raleigh, North Carolina, USA

Min Chi
mchi@ncsu.edu
North Carolina State University
Raleigh, North Carolina, USA

ABSTRACT

We explore eXplainable AI (XAI) to enhance user experience and understand the value of explanations in AI-driven pedagogical decisions within an Intelligent Pedagogical Agent (IPA). Our real-time and personalized explanations cater to students' attitudes to promote learning. In our empirical study, we evaluate the effectiveness of personalized explanations by comparing three versions of the IPA: (1) personalized explanations and suggestions, (2) suggestions but no explanations, and (3) no suggestions. Our results show the IPA with personalized explanations significantly improves students' learning outcomes compared to the other versions.

CCS CONCEPTS

• Human-centered computing → Empirical studies in HCI; **• Computing methodologies → Inverse reinforcement learning**; **• Applied computing → Computer-assisted instruction**; **E-learning**; **Interactive learning environments**.

ACM Reference Format:
John Wesley Hostetter, Cristina Conati, Xi Yang, Mark Abdelshiheed, Tiffany Barnes, and Min Chi. 2023. XAI to Increase the Effectiveness of an Intelligent Pedagogical Agent. In *ACM International Conference on Intelligent Virtual Agents (IVA '23), September 19–22, 2023, Würzburg, Germany.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3570945.3607301

1 INTRODUCTION

An important goal of eXplainable AI (XAI) is to make AI systems more transparent and trustworthy by displaying their inner workings. While some research focuses on increasing the interpretability of black-box algorithms for model validation and improvement, there has been a growing interest in XAI to improve the *end-users' experience* with AI applications. Studies have found the importance of AI explanations for end-users [32, 53], but that explanations aren't always wanted or beneficial [13, 20, 57]. There is a general agreement that the need for explanations depends on the context of the AI application or task criticality [13, 44] and that user differences like cognitive abilities and personality traits play a role in determining *if*, *when*, and *how* explanations are valuable [16, 31, 41, 43, 45, 54]. These findings call for research on personalized XAI to understand how to create AI systems that effectively explain their actions and decisions to the right users at the right time.

This led us to explore XAI to enhance *student's* experience and understand explanations' value in AI-driven pedagogical decisions within an Intelligent Pedagogical Agent (IPA). This work has two main contributions: **first**, is that we explore the personalization of explanations to user traits that have not been studied before in XAI —*students' attitude toward learning* within an IPA (which we will refer to as *learning attitude* from now on). Our **second** contribution is that, to the best of our knowledge, this is the first *empirical study exploring the effectiveness of delivering real-time personalized explanations*. Previous studies have shown the need for personalization by providing non-personalized explanations to users and analyzing their perceptions based on different traits [16, 31, 41, 45, 54]. Some XAI research has begun to address this need, such as designing personalized explanations for a music recommender system [38], tailored to traits identified as relevant in prior studies [41, 42]. However, these personalized explanations were only tested on users with higher or lower levels of the targeted traits rather than in real-time interactions. In contrast, our system predicts users' learning attitudes in real-time and personalizes the explanations of its individualized pedagogical interventions for each user.

To investigate the effectiveness of personalizing explanations for a pedagogical action/decision to student learning attitude, we extend an existing IPA —called *Pyrenees* [1, 3–5], that helps students make decisions on how to learn from a pool of available probability problems [26, 27]. In Pyrenees, students first make some pedagogical decisions on whether to solve the next problem by themselves, in collaboration with the IPA, or look at the problem solution as a worked-out example. Although the student makes the decisions, Pyrenees uses Hierarchical Reinforcement Learning [53] or expert-designed rules to offer suggestions on the decision based on how the current student is progressing through the available problems.

Our personalized explanations are generated based on students' attitudes toward learning since prior research has shown that this plays a crucial role in students' motivation [2, 7, 8], engagement in class [11], confidence [33], and even perception of or retention

IVA '23, September 19–22, 2023, Würzburg, Germany

ACM ISBN 978-1-4503-9994-4/23/09…$15.00
https://doi.org/10.1145/3570945.3607301

of the material [47]. For example, an individual's attitude toward learning is essential when it comes to determining how well he or she will perform in math [36, 47], physics [23], or other scientific fields, such as microbiology [18]. Here, we define learning attitudes by extending our prior work where Inverse Reinforcement Learning (IRL) was applied to student-IPA logs to infer learning intentions [58]. We further extended Pyrenees by incorporating a deep learning classifier to predict students' learning attitudes in real time, generating personalized explanations with our predictions.

We evaluate personalized explanations in a formal user study comparing an IPA that *intervenes and explains* ***(Intervene-Explain)*** against one that *intervenes without explaining* ***(Intervene-Only)***, and *one with no intervention* ***(StuChoice)***. Results show that personalized explanations significantly improve students' learning gains over the other two, demonstrating their crucial role in improving learning. The paper is organized as follows: Section 2 describes related work; Sections 3, 4 and 5 presents Pyrenees as well as the real-time classifier and personalized explanations; Section 6 describes the experiment setup; Section 7 shows our results and 8 present results and conclusion, ethics, and impact of our work.

2 RELATED WORK

Much of the research on XAI for end-users is in the field of recommender systems. Studies have shown that individual differences, such as users' levels of neuroticism and decision-making style, can impact preferences for different types of explanations [31]. In online dating platforms, men tend to focus on their preferences more than women; users with "choosy posts" benefit more from reciprocal explanations [30]. Research also shows that personality affects preferences for movie recommendations and corresponding explanations [10]. In the context of a music recommender system, the need for cognition (a trait that measures one's appreciation for effortful cognitive activities [14]) [41] as well as musical sophistication and openness (one of the personality traits in the Big-Five Factor Model [21], which measures the breadth and complexity of an individual's mental and experiential life) [42] have a significant impact on explanation effectiveness.

Beyond recommender systems, explanations of AI agents' suggestions for playing online games are only valid for users with low game ability [54]. Research has investigated the impact of individual differences such as the need for cognition, reading proficiency, and conscientiousness (another personality trait in the Big-Five Factor Model [21]) on the effectiveness of explanations for AI-driven hints generated by IPAs [16]. Most evidence generated is not personalized, and minimal results exist on the effectiveness of personalizing AI explanations to user differences. Some papers mention personalized explanations but refer to explaining the actions of the AI rather than the accompanying explanations [10, 31, 56]. In contrast, in our work, the explanations are personalized; the same system action/decision may be explained differently to different users.

The Open Learner Model (OLM) [15] increases IPA transparency by showing students a current assessment of their abilities. Although there is evidence of OLM improving learning [35, 49], self-perception [49], and facilitating more trust between the student and the IPA [37], it remains unclear how this might be applied to enhance the interpretability and explainability of IPA. An attempt was made by [9], where explanations were added to an OLM, but they are essentially textual rephrases of the OLM assessment.

Closer to our objectives, the work by [16] and [53] explored providing explanations for IPA decisions derived by Reinforcement Learning. The explanations given to the students were short sentences focused on the high-level benefits of the activity. The study offered preliminary evidence showing students saved time and gained more autonomy within the IPA. This suggests that students may benefit from knowing the high-level motivation regarding pedagogical decisions [51]. However, neither of these works provides personalized explanations [16, 53]. In this work, explanations go beyond OLM or high-level motivations to convey a more fine-grained understanding of why the IPA's suggestion is helpful for that particular student and how the underlying AI derived the decision.

3 PYRENEES

Pyrenees is a web-based IPA that teaches ten probability principles (e.g., Addition Theorem and Bayes' Theorem) [3, 4, 6]. During training, twelve problems are shown in a fixed sequence, allowing students to practice by choosing to solve it alone (PS), solving it *collaboratively* with the IPA (CPS), or seeing the solution as a worked-out example (WE). Students work through the problem step-by-step, define variables, type equations, etc. If PS is chosen, the IPA asks questions to elicit the next step's answer from the student; if CPS is chosen, the IPA chooses to elicit or tell the answer; if WE are chosen, the IPA tells the answer. The IPA provides feedback with a short message or hints if the student gives a wrong answer. Hints can also be requested by clicking a button, which is organized in an increasingly specific order, where the last message shows the student exactly what to do.

To prevent decision fatigue [48], students make 10 problem-level decisions on whether to present the next problem as a WE, a PS, or CPS. The IPA challenges students' decisions at selected times and presents them with a choice. This decision process is based on an Expert-designed policy, or Deep RL [53]. Our work uses offline, off-policy Deep Hierarchical RL to induce policies from a historical dataset containing 1,148 students' interaction logs collected over 6 semesters using the same IPA, procedure, materials, and problems. The components for RL induction are defined as follows:

State: From the student-system interaction logs, 142 features were extracted to represent the student learning state, which can be categorized into five groups: **Autonomy** (10 features): the amount of work done by a student, such as several elicits since the last tell; **Temporal** (29 features): time-related information about the student's behavior, such as the average time per step; **Problem-Solving** (35 features): information about the current problem-solving context, such as problem difficulty; **Performance** (57 features): information about the student's performance so far, such as the percentage of correct entries; **Hints** (11 features): information about the student's hint usage, such as the total number of hints requested.

Action: Our IPA makes problem and step-level decisions; there are two possible actions at the step level (elicit/tell) and three at the problem level (WE/PS/CPS).

Reward: We used Normalized Learning Gain (NLG) as a delayed reward, which measures students' learning gain regardless of their initial competence [1, 4, 7, 26, 27]. NLG is calculated as $\frac{posttest-pretest}{\sqrt{1-pretest}}$, where the max score is 1 for both the pre-test and post-test. Our Deep RL aims to induce policies for maximizing learning gains.

The original explanations in Pyrenees were designed to convey the benefit of taking **the suggested pedagogical actions**, based on research in learning science and cognitive science. For example, viewing examples is beneficial for learning new content [39, 40], and thus an explanation for WE would state "The AI agent thinks you should view this problem as a Worked Example to learn how some new rules work." Furthermore, in our original explanations, a simple action-based explanation for another WE would state: "The AI agent thinks you would benefit from viewing this problem as a Worked Example to save time." Similarly, if the policy decision was that the following problem should be a PS, the message would state something like: "The AI agent thinks you should solve this problem yourself to improve learning." In this work, we expanded these explanations by personalizing them to the student's learning attitude in an effort to further improve their learning. We introduce our real-time learning attitude classifier in the following Section 4.

4 LEARNING ATTITUDE CLASSIFIER

Our personalized explanations were tailored to each student's learning attitudes. To infer their attitudes, we leverage our prior **Inverse Reinforcement Learning (IRL)** research in [58]. Specifically, the central idea is to frame the problem of determining students' learning attitudes as inferring their reward function from their demonstrated behavior or decision trajectories (i.e., their pedagogical decision-making history) through IRL.

To do so, we formalize the student's sequential decision process as a Markov Decision Process (MDP). An MDP describes a stochastic control process using a tuple $< S, A, R, T, \gamma >$. In the context of IPA, states S are often represented by vectors composed of relevant learning environment features, such as the percentage of correct attempts a student has made so far and so on. Actions A are the possible pedagogical decisions, such as WE, PS, or CPS. The transition probability T can be estimated from training data. $\gamma \in [0, 1)$ denotes a discount factor for future rewards. Since we don't know the student's reward function R, the task of IRL can be described as a stochastic control process where we have $MDP \backslash R =< S, A, T, \gamma >$ together with some demonstrated trajectories $\mathcal{T}$. More specifically, the students' demonstrated decision trajectories can be represented as $s_1 \longrightarrow a_1 s_2 \longrightarrow a_2 \cdots s_n \longrightarrow a_n$. Here $s_i \longrightarrow a_i s_{i+1}$ indicates that at the i^{th} moment, the student was in some learning state called s_i, a pedagogical decision referred to as a_i was carried out and that led the student to be in the new learning state represented by s_{i+1}.

We apply IRL to their decision-making trajectories to infer students' learning attitudes to learn their reward function. Formally, we denote the input N demonstrated trajectories as $\mathcal{T} = \{\xi_1, ..., \xi_N\}$ and each trajectory is composed of a set of state-action pairs: $\xi_i = \{(s_1, a_1), (s_2, a_2), ...\}$. *The assumption here is that students would make pedagogical decisions to maximize their expected long-term reward function.* Once the reward function is learned, any RL or Deep RL method can further induce the strategy followed by $\mathcal{T}$. Typically, IRL approaches are designed to model data assuming all trajectories have a single reward function. However, students may have different learning attitudes; some may want to finish the training on the IPA as fast as possible, while others may wish to learn as much as possible. To account for this heterogeneity, we employed *Expectation-Maximization IRL (EM-IRL)* proposed in [58] to extract heterogeneous learning attitudes.

Our training dataset consists of 127 students who each spent an average of 2 hours on Pyrenees and completed around 400 steps, using the same IPA and following the same procedures, materials, and problems. **States:** The same 142 state features were leveraged, and K-means clustering was applied to each feature category to generate discrete states. The number of states for each category was determined by selecting the elbow of errors in the clustering results: Autonomy (3), Temporal (4), Problem-Solving (3), Performance (4), and Hints (3), resulting in 432 discrete states in total. Transition probabilities were estimated based on these discrete states using all available data. **Actions:** Students can make pedagogical decisions at both the step and problem level, such as problem-solving or worked-out examples. By using EM-IRL on the training dataset, three types of learning attitudes were identified: (1) **Learning-oriented:** students learn as much material as is possible regardless of the time that is spent; (2) **Efficient-oriented:** students efficiently learn significant amounts of the material but spend less time than their "Learning-oriented" peers; (3) **No-learning:** students spent less time and also failed to learn the material.

We leverage Long Short-Term Memory (LSTM) [25] to predict students' learning attitudes. LSTM is capable of memorizing temporal dependencies over a long period and has shown extensive prospects in a variety of sequential labeling applications, such as climate changes, healthcare, and traffic monitoring [34][29][28]. This work uses LSTMs to incorporate the previous hidden state as input for each subsequent step; this mechanism allows for accumulating memory and modeling dynamic information, specifically capturing student learning events over time.

Since students are likely to change their learning attitudes while training, we developed an online LSTM-based early prediction model to classify students' learning attitudes in real-time on each problem. For the m^{th} problem, our LSTM classifier will leverage trajectories of the student's interactions with the system up to the end of the $(m-1)^{th}$ problem as input to predict their learning attitude before they make a pedagogical decision. Note that for all students, no decision was made for the first ($m = 1$) and 8th ($m = 8$) problems; thus, all students make ten pedagogical decisions.

As shown in Figure 1, 5-fold cross-validation results found that by only using decision-making history from the first problem, the early prediction accuracy was 0.67 and the AUC was 0.77 for $m = 2$. The LSTM's early prediction performance on the ten problems improved with an accuracy of 0.79 and AUC of 0.89 for $m = 3$. The AUC performance was above 0.89 until the last problem, $m = 12$. Thus, we use LSTM models to predict their learning attitude starting from $m = 3$ to ensure accurate early predictions. In subsequent problems, our classifier relies on the students' interactive logs to predict their learning attitude. If an intervention is required, personalized explanations will be provided based on their predicted learning attitude. By conducting the early prediction before each problem, the student's attitude towards learning can be monitored so that the personalized explanations provided remain relevant to the student. The inspirations behind the design of our personalized explanations are shown next in Section 5.

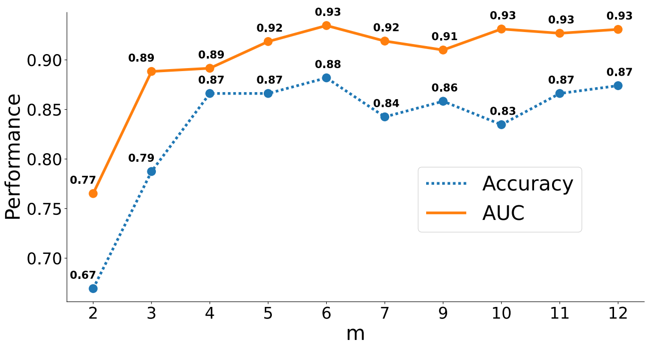


Figure 1: 5-fold cross-validation results plotted against the number of past problems used in the learning classifier's input. As the number of historical data increases, the classifier's performance improves in predicting the student's learning attitude. Based on this plot, we used interaction data from the past three problems to predict the learning attitude to achieve early prediction and performance.

5 EXPLAINING THE IPA

5.1 What is an Explanation and How to Explain

People have *social expectations* when they evaluate the quality of an explanation [24]. Evidence suggests people anthropomorphize artificial agents [19] and naturally expect explanations given by artificial agents to comply with our social norms on what constitutes a quality explanation [44]. We drew inspiration from the social sciences to develop our personalized explanations by following Miller's work [44]. According to Miller, explanations are expected to be *contrastive* in that they employ *counterfactual cases* [44]. In other words, people do not simply ask *why* some event happened, but rather, they ask *why* that occurred *instead of* some other event. Research exploring the need for XAI in IPAs supports students' desire for explanations answering *why* more often than *how* [50].

In our framework, explanations are provided to a student when the IPA intervenes. Therefore, our IPA's explanations are written where the *counterfactual* or the *instead of* is assumed to be the student's choice, and the explanation is answering *why* the student should follow the IPA's pedagogical intervention. It should be noted our explanations are *selected* —meaning they do not contain *all* causes as humans rarely ever find this beneficial [55]; further, they *do not contain probabilities* as evidence suggests probabilities or statistical relationships are unfulfilling [44]. In essence, our explanations are in the form of "Why did event A happen instead of event B?", which requires an explanation appropriate for this *why*-question. Miller states these questions are some of the most challenging to produce explanations for, as they require counterfactual, associative, and interventionist reasoning [44].

5.2 Designing Our Explanations

To answer this *why*-question in the IPA setting, we take a step towards personalizing explanations according to the student's attitude towards learning, and to the best of our knowledge, there is no existing published research outlining such an initiative. Although providing explanations to users can be a "double-edged sword" [20] —as they are sometimes unwanted or not beneficial [12, 13, 20] —it remains imperative that pedagogical interventions and explanations are highly individualized and personalized to a student's need since a negative impact on students' learning could potentially be long-lasting [16]. Furthermore, in light of the evidence that suggests XAI in IPAs may help to promote trust in the system [17] and may achieve an overall positive effect on a student's learning [16], we propose to *personalize* our explanations according to the student's attitude towards learning to *improve their learning gain*. However, since interfering with the student's autonomy (i.e., the student's pedagogical decision) may be an *autonomy-suppressive* teaching behavior according to Self-Determination Theory [8], personalized explanations should possess *intention* behind *why* we are interfering with the student's decision —hence our prediction of their learning attitude; in this sense, our personalized explanations promote an *autonomy-supportive* teacher behavior by *fostering relevance* to the individual student. These personalized explanations were written per Overton's definition for a *why*-question explanation [46], which consists of two components: (1) the *explanans* and (2) the *explanandum*. The explanans is an answer to the *why*-question whereas the explanandum is the *presupposition* (e.g., the givens or context within which the explanation is contained). Thus, the general structure of our explanans is: <*what the student should do*> and <*why*> where the student's choice or learning attitude are examples of the explanandum or presupposition.

5.3 Personalizing Explanations to Students

In Pyrenees , there are two interaction levels: (1) problem-level; (2) step-level. At the problem level, Pyrenees will show the student a problem description and ask the student if they want: (1) to review a worked example (WE), (2) solve it alone (PS), or (3) *to collaborate* on the problem together (CPS). The second level of decision-making, called step-level, occurs only if the student chose to *collaborate* on a given problem. A step is a brief amount of work done to solve the problem, such as applying a probability principle. If the student belongs to an *interactive* policy, the IPA will intervene if the student makes a sub-optimal pedagogical decision during a critical moment. The personalized explanation is given based on the student's initial decision, the suggested decision of the IPA, and the student's learning attitude as predicted by the LSTM classifier. Table 1 provides examples of such personalized explanations, detailing the student's initial decision (1^{st} column), the IPA's suggested approach (2^{nd} column), and the predicted student learning attitude (3^{rd} column).

6 EXPERIMENT SETUP

Participants: Pyrenees was given to students as a homework assignment in an undergraduate Computer Science class in the Fall of 2021. Students were told to complete the study in one week and will be graded based on demonstrated effort rather than learning performance. 180 students were *randomly assigned* into three conditions: Intervene-Only ($N = 74$), Intervene-Explain ($N = 43$) and StuChoice ($N = 63$). *It is important to note that the difference in size among the conditions is because we prioritized having a sufficient number of participants in the Intervene-Only and StuChoice conditions to perform a meaningful analysis.* Due to preparation for final exams and the length of study, 151 students completed the

Table 1: A small sample of personalized explanations regarding student's & IPA's choice, and the student's learning attitude on problem-level decision-making. Bold text shows how the explanation is tailored to the student's learning attitude.

Student	IPA	Attitude	Explanation
WE	**PS**	Learning-oriented	"You should solve this problem by yourself to **reinforce your knowledge** since it is a good exercise for applying what you have learned."
		Efficient-oriented	"You should solve this problem yourself, as the difficulty level of this problem is just right for you. **By doing so, you will learn more efficiently.**"
CPS	**PS**	Learning-oriented	"Make sure you solve this problem by yourself to **further consolidate your knowledge**, as this problem is at the right level of difficulty for your current knowledge."
		Efficient-oriented	"This one you should solve yourself since you need to practice your knowledge so that **you can complete the training efficiently**."
WE	**CPS**	No-learning	"We will solve this problem together **since we are good on time and you will learn more this way.**"
		Learning-oriented	"This time, let us work together to solve the problem **to improve your learning outcome** given how far you have come and the problem's difficulty."
		Efficient-oriented	"Let's solve this problem together based on your performance so far and its difficulty level **so that you can learn more effectively later**."
PS	**WE**	Learning-oriented	"Let me show you how to solve this problem because my solution will be more efficient and it will **benefit your learning outcomes."**
		Efficient-oriented	"Let me show you how to solve the next problem (based on the time you have already spent so far) that will **make your training experience efficient and on time**. "

study; the completion rate between conditions was insignificant according to the Chi-square test: $\chi^2(2, 180) = 3.2929$, $p = .193$.

Experiment procedure & grading: The entire experiment procedure is given as a homework assignment and must be completed independently. It consists of four stages following the strict order:

(1) **Textbook:** Students read about the ten probability principles and review examples. The textbook is shown online.
(2) **Pre-test:** Students' took a pretest consisting of fourteen problems: ten single-principle problems, one for each probability theorem being taught, and four multiple-principle problems. Each problem requires a detailed step-by-step solution, with the final answer provided separately. No feedback was given to their answers, students were not allowed to return to earlier problems, and the textbook was unavailable during the pretest; the same applied to the post-test.
(3) **Training with the IPA:** Students worked through the same twelve problems in the same order with Pyrenees . The steps to solve each problem ranged from twenty to fifty and included variable definitions, principle applications, and equation solving. The experimental conditions only differed in intervention with/out personalized explanations.
(4) **Post-test:** Contains fourteen isomorphic problems (ten single and four multiple-principle) from the pre-test and six new multiple-principle problems designed to be significantly more challenging.

All tests were double-blind graded by two experienced graders, and normalized to $[0, 1]$ for comparison purposes.

7 RESULTS

7.1 Learning Performance

We evaluate the effectiveness of the personalized interventions in both learning performance and time on task (namely, time students spent working through the twelve problems in Pyrenees) because there is often a trade-off between these measures. As introduced earlier in Section 3, learning performance is measured with Normalized Learning Gain (NLG), which measures a student's learning gain irrespective of their incoming competence.

NLG is defined as $\frac{posttest-pretest}{\sqrt{1-pretest}}$, where 1 is the maximum score for both pre- and post-test. We compute two different types of learning gains, one that involves the pretest and only the isomorphic questions in the post-test (*isomorphic* [*or Iso*] *NLG* from now on) and one that involves the pre-test and all the questions in the post-tests, including the six much more difficult ones (*NLG* from now on). Time on task was derived from the Pyrenees logs, where every click students made on the interface included a timestamp. These interactive student-system logs were used for our Deep RL policy and real-time learning attitude prediction.

For our subsequent statistical analysis, six students were excluded (four from StuChoice and two from Intervene-Explain) due to receiving a perfect performance in the pretest to accommodate for the ceiling effect of our learning metrics. Additionally, three students with isomorphic NLG scores greater than three standard deviations from the mean were removed (one from StuChoice and two from Intervene-Only), as they represent extreme outlier students who did not pay attention to or engage with the IPA. The final group sizes were Intervene-Only ($N = 56$), Intervene-Explain ($N = 39$) and StuChoice ($N = 49$).

Table 2 shows the mean and standard deviation (SD) of students' learning gains as well as time on task in hours (h). We find no significant differences among the conditions' pretest scores (the 2nd column): *Welch's* $F(2, 91.426) = .364, p = .696$, indicating that students started with similar knowledge. We then compare the learning gains of the three conditions. According to Shapiro-Wilk's test, Iso NLGs and NLG scores were not normally distributed, and

Table 2: Mean (SD) of learning performance and time on task.

Condition	Pre	Iso Post	Post	Iso NLG	NLG	Time (h)
Intervene-Explain	.731 (.149)	.814 (.149)	.729 (.175)	.136 (.271)	-.037 (.316)	1.659 (.467)
Intervene-Only	.767 (.185)	.792 (.149)	.701 (.160)	-.065 (.453)	-.290 (.529)	1.776 (.925)
StuChoice	.753 (.195)	.763 (.201)	.689 (.217)	-.024 (.328)	-.193 (.362)	1.933 (.817)

concerning Levene's test for equality of variances, variances were homogeneous. Since the normality assumption is invalid and sample sizes are unequal, a non-parametric statistical test —the Kruskal-Wallis H test —was conducted to compare the medians as it does not assume an underlying normal distribution. If H test results were significant, pairwise comparisons were performed using Dunn's procedure with Bonferroni adjustments to control for Type I errors.
Comparison of Iso-NLG Scores: A significant difference was found in Iso NLGs' medians across conditions: *H test's* $\chi^2(2) = 8.731$, $p = .013$. Dunn's procedure with Bonferroni adjustments revealed Intervene-Explain significantly outperformed StuChoice ($p = .023$) as well as Intervene-Only ($p = .031$), where the difference between the latter two was insignificant ($p = 1.0$).

These results suggest personalized explanations improved students' ability to learn principles needed to solve similar problems in the pre-test and post-test using Pyrenees . The students in Intervene-Explain were the only ones that improved from the pre-test to the post-test, as verified by a pair-sampled t-test ($t(38) = 3.834$, $p < .001$). In contrast, there was no statistically significant improvement from the pre-test to the post-test for the other two: $t(48) = .436$, $p = .665$ for StuChoice, and $t(53) = 1.550$, $p = .127$ for Intervene-Only.
Comparison of NLG Scores: Median NLGs across the three conditions differed significantly: *H test's* $\chi^2(2) = 7.244$, $p = .027$. Dunn's procedure with Bonferroni adjustments found significant differences between Intervene-Explain and Intervene-Only ($p = .039$) and a marginal difference between Intervene-Explain and StuChoice ($p = .071$). The difference in medians between StuChoice and Intervene-Only is insignificant ($p = 1.0$). These results suggest that personalized explanations were compelling even in complex problems, with students performing better than those in the Intervene-Only condition. Although the performance difference between the explanation and no-intervention conditions is insignificant, the trend still favors the explanation condition.

To summarize, explanations personalized to a student's learning attitude enhanced student learning compared to the StuChoice or Intervene-Only conditions. Notably, this increase in learning is not related to students taking more time to solve the twelve practice problems with the IPA since we found no significant difference in time of task across the conditions (last column of Table 2) per a Welch's ANOVA: *Welch's* $F(2, 91.022) = 1.983, p = .144$, with a trend showing students in the Intervene-Explain condition taking the least amount of time.

7.2 Personality Traits Analysis

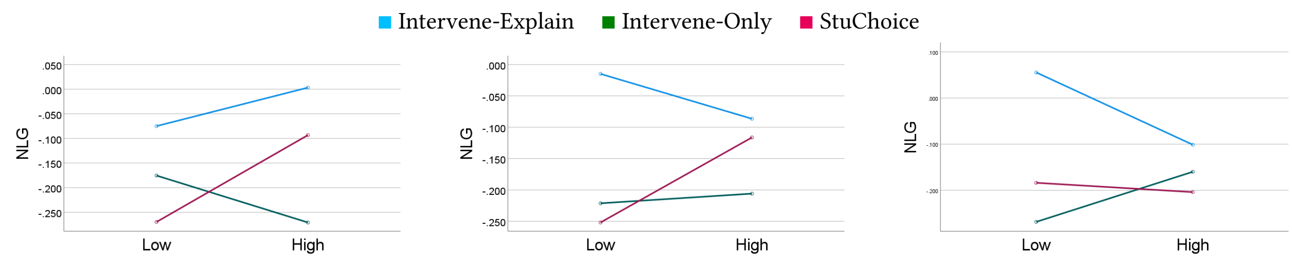
Here, we investigate whether the effectiveness of personalized explanations delivered by Pyrenees might be modulated by any personality traits in the Big-Five personality model [21]. Table 3 lists the five traits and their definition. The measures for these traits were obtained by asking our study participants to fill out an optional Ten-Item Personality Inventory (TIPI) survey afterward [22]. The response rate for our optional survey was 95.83%, with only three students from StuChoice and three students from Intervene-Only choosing not to participate, leaving us with Intervene-Explain ($N = 39$), Intervene-Only ($N = 53$), and StuChoice ($N = 46$). A one-way ANCOVA was performed for each personality trait, with that trait as a co-variate, experimental condition as a factor, and NLG as the dependent variable. The ANCOVAs returned significant interaction effects between conditions and the personality traits: *extraversion* ($F(2, 134) = 3.111$, $p = .048$, partial $\eta^2 = .044$), *agreeableness* ($F(2, 134) = 3.305$, $p = .040$, partial $\eta^2 = .047$) and *neuroticism* ($F(2, 134) = 3.103$, $p = .048$, partial $\eta^2 = .044$). These interaction effects are illustrated in Figure 2, where users are divided into low and high groups for each trait based on a median split of their corresponding TIPI values. Visual inspection of Figure 2 confirms that, for all three personality traits, students in Intervene-Explain show higher NLGs overall than the others. Pairwise comparisons reveal interesting trends summarized in Table 4, indicating the significant impact of personalized explanations on students with different levels of the three personality traits, as measured by Cohen's d effect size.

(a) **Extraversion:** Figure 2a shows students' NLGs for Low vs. High extraversion across the three conditions. Pairwise comparisons show a significant difference only within the High extraversion group, where Intervene-Explain significantly outperforms Intervene-Only ($p = .018$ and $d = 0.739$), with no significant difference between the StuChoice and Intervene-Only.
(b) **Agreeableness:** Figure 2b shows students' NLGs for Low vs. High agreeableness across the three conditions. Pairwise comparisons show a significant difference only within the Low Agreeableness group, where Intervene-Explain significantly outperforms both StuChoice ($p = .019$ and $d = 0.714$) and Intervene-Only ($p = .043$ and $d = 0.652$), with no significant difference between these latter two.
(c) **Neuroticism:** In Figure 2c shows students' NLGs for Low vs. High neuroticism across the three conditions. Pairwise comparisons show a significant difference only within the Low neuroticism group, where Intervene-Explain significantly outperformed both Intervene-Only ($N = 26$) and StuChoice ($N = 25$), $p = .006$ ($d = 0.964$) and $p = .041$ ($d = .770$).

In terms of implication for further personalization of the Pyrenees explanations tailored to student learning attitude, these results indicate that the current personalized explanations are most effective for students with Low agreeableness, Low Neuroticism, and High Extroversion. Since they don't seem to harm the other groups, one could say that there is no need to consider student personality

Table 3: Five user characteristics and their definitions according to the "Big Five" personality dimensions [52].

User Characteristic	Definition
Agreeableness	An agreeable person is fundamentally generous, sympathetic to others, and eager to help them.
Conscientiousness	Refers to self-control and the active process of planning, organizing, and carrying out tasks.
Extraversion	Includes traits such as sociability, assertiveness, activity, and talkativeness.
Neuroticism	General tendency to experience negative effects such as fear, sadness, embarrassment, anger, guilt, and disgust.
Openness	Openness to Experience includes active imagination, aesthetic sensitivity, attentiveness to inner feelings, a preference for variety, intellectual curiosity, and independence of judgment.



(a) Extraversion **(b) Agreeableness** **(c) Neuroticism**

Figure 2: Interaction effect between level of user characteristic (low and high) and condition on students' NLG.

Table 4: The most effective pedagogical strategy for students based on their identified personalities, preferring Student Choice when no significant differences are found.

Personality Trait	Group	Comparison	Conclusion
Extraversion	Low	No significance	-
	High	Intervene-Explain > Intervene-Only	Do not provide Intervene-Only
Agreeableness	Low	Intervene-Explain > Intervene-Only, StuChoice	Provide Intervene & Explain
	High	No significance	-
Neuroticism	Low	Intervene-Explain > StuChoice, Intervene-Only	Provide Intervene & Explain
	High	No significance	-

when delivering the attitude-based explanations; they should be delivered to everyone since they help some students and don't hurt others. However, this hypothesis should be verified empirically with further studies and more data to ensure that the lack of significant effects in some comparisons is not due to a lack of power.

8 DISCUSSION, ETHICS & BROADER IMPACTS

This work investigates the effectiveness of providing real-time personalized explanations to students based on their learning attitudes. Using an EM-IRL algorithm to determine learning attitudes and an LSTM model for real-time predictions, our personalized explanations are generated based on the student's decision, the IPA's suggestion, and the predicted attitude. Results show that Intervene-Explain outperforms Intervene-Only and StuChoice in terms of Iso NLG and NLG. Our findings suggest that providing personalized explanations according to students' attitudes toward learning can enhance learning and generalization. The data used in this study were anonymously obtained through an IRB-approved protocol and scored against test cases. No demographic data or grades were collected. This research focuses on providing XAI for end-users; it is worth noting that interpreting XAI is not always clear, and the models used here may not always be accurate.

9 ACKNOWLEDGMENT

This research was supported by the NSF Grants: Integrated Data-driven Technologies for Individualized Instruction in STEM Learning Environments (#1726550), CAREER: Improving Adaptive Decision Making in Interactive Learning Environments (#1651909), and Generalizing Data-Driven Technologies to Improve Individualized STEM Instruction by Intelligent Tutors (#2013502), and the AI Institute for Engaged Learning (DRL-#2112635). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Mark Abdelshiheed. 2023. *Combining Reinforcement Learning and Three Learning Theories to Achieve Transfer and Bridge Metacognitive Knowledge Gap.* North Carolina State University.

[2] Mark Abdelshiheed, Tiffany Barnes, and Min Chi. 2023. How and When: The Impact of Metacognitive Knowledge Instruction and Motivation on Transfer across Intelligent Tutoring Systems. *International Journal of Artificial Intelligence in Education* (2023).

[3] Mark Abdelshiheed, John Wesley Hostetter, Tiffany Barnes, and Min Chi. 2023. Bridging Declarative, Procedural, and Conditional Metacognitive Knowledge Gap Using Deep Reinforcement Learning. In *Proceedings of the 45th annual conference of the cognitive science society.*

[4] Mark Abdelshiheed, John Wesley Hostetter, Tiffany Barnes, and Min Chi. 2023. Leveraging Deep Reinforcement Learning for Metacognitive Interventions across Intelligent Tutoring Systems. In *Proceedings of the 24th International Conference on Artificial Intelligence in Education.* Springer International Publishing.

[5] Mark Abdelshiheed, John Wesley Hostetter, Preya Shabrina, Tiffany Barnes, and Min Chi. 2022. The Power of Nudging: Exploring Three Interventions for Metacognitive Skills Instruction across Intelligent Tutoring Systems. In *CogSci*, Vol. 44.

[6] Mark Abdelshiheed, John Wesley Hostetter, Xi Yang, Tiffany Barnes, and Min Chi. 2022. Mixing Backward-with Forward-Chaining for Metacognitive Skill Acquisition and Transfer. In *AIED*. Springer, 546–552.

[7] Mark Abdelshiheed, Guojing Zhou, Mehak Maniktala, Tiffany Barnes, and Min Chi. 2020. Metacognition and Motivation: The Role of Time-Awareness in Preparation for Future Learning. In *CogSci*. 945–951.

[8] Avi Assor, Haya Kaplan, and Guy Roth. 2002. *Choice is good, but relevance is excellent: Autonomy-enhancing and suppressing teacher behaviours predicting students' engagement in schoolwork.* Vol. 72. 261–278. Issue 2.

[9] Jordan Barria-Pineda et al. 2019. Explaining Need-Based Educational Recommendations Using Interactive Open Learner Models. In *Adjunct Publication of the 27th Conference on UMAP* (Larnaca, Cyprus). ACM, New York, NY, USA, 273–277.

[10] Shlomo Berkovsky, Ronnie Taib, and Dan Conway. 2017. How to Recommend? User Trust Factors in Movie Recommender Systems. In *Proceedings of the 22nd International Conference on IUI* (Limassol, Cyprus). ACM, New York, NY, USA, 287–300.

[11] Colin Bryson and Len Hand. 2007. The role of engagement in inspiring teaching and learning. *Innovations in Education and Teaching International* 44, 4 (2007), 349–362.

[12] Andrea Bunt et al. 2007. Understanding the Utility of Rationale in a Mixed-Initiative System for GUI Customization. In *User Modeling 2007*. Springer Berlin Heidelberg, Berlin, Heidelberg, 147–156.

[13] Andrea Bunt, Matthew Lount, and Catherine Lauzon. 2012. Are Explanations Always Important? A Study of Deployed, Low-Cost Intelligent Interactive Systems. In *Proceedings of the 2012 ACM International Conference on IUI* (Lisbon, Portugal). ACM, New York, NY, USA, 169–178.

[14] John T. Cacioppo, Richard E. Petty, and Chuan Feng Kao. 1984. The Efficient Assessment of Need for Cognition. *Journal of Personality Assessment* 48, 3 (1984), 306–307.

[15] William J. Clancey. 1986. From Guidon to Neomycin and Heracles in Twenty Short Lessons. *AI Magazine* 7, 3 (Jul. 1986), 40.

[16] Cristina Conati et al. 2021. Toward personalized XAI: A case study in intelligent tutoring systems. *Artificial Intelligence* 298 (2021), 103503.

[17] Cristina Conati, Kaska Porayska-Pomsta, and Manolis Mavrikis. 2018. AI in Education needs interpretable machine learning: Lessons from Open Learner Modelling. *CoRR* abs/1807.00154 (2018). arXiv:1807.00154 http://arxiv.org/abs/1807.00154

[18] Dana E. Craker and Abdulaziz Elfessi. 2006. Attitudes Toward Science of Students Enrolled in Introductory Level Science Courses at UW-La Crosse.

[19] Maartje de Graaf and Bertram F. Malle. 2017. How People Explain Action (and Autonomous Intelligent Systems Should Too). In *AAAI Fall Symposia*.

[20] Kate Ehrlich et al. 2011. Taking Advice from Intelligent Systems: The Double-Edged Sword of Explanations *(IUI '11)*. ACM, New York, NY, USA, 125–134.

[21] L R Goldberg. 1990. An alternative "description of personality": the big-five factor structure. *J Pers Soc Psychol* 59, 6 (Dec. 1990), 1216–1229.

[22] Samuel D Gosling, Peter J Rentfrow, and William B Swann. 2003. A very brief measure of the Big-Five personality domains. *Journal of Research in Personality* 37, 6 (2003), 504–528.

[23] Ryan Manuel D. Guido. 2018. Attitude and Motivation towards Learning Physics.

[24] Denis J. Hilton. 1990. Conversational processes and causal explanation. *Psychological Bulletin* 107 (1990), 65–81. Issue 1.

[25] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[26] John Wesley Hostetter, Mark Abdelshiheed, Tiffany Barnes, and Min Chi. 2023. Leveraging Fuzzy Logic Towards More Explainable Reinforcement Learning-Induced Pedagogical Policies on Intelligent Tutoring Systems. In *2023 IEEE International Conference on Fuzzy Systems*. IEEE.

[27] John Wesley Hostetter, Mark Abdelshiheed, Tiffany Barnes, and Min Chi. 2023. A Self-Organizing Neuro-Fuzzy Q-Network: Systematic Design with Offline Hybrid Learning. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS).*

[28] Xiaowei Jia, Ankush Khandelwal, Guruprasad Nayak, James Gerber, Kimberly Carlson, Paul West, and Vipin Kumar. 2017. Incremental Dual-memory LSTM in Land Cover Prediction. In *SIGKDD*. ACM.

[29] Xiaowei Jia, Ankush Khandelwal, Guruprasad Nayak, James Gerber, Kimberly Carlson, Paul West, and Vipin Kumar. 2017. Predict land covers with transition modeling and incremental learning. In *Proceedings of the 2017 SDM.*

[30] Akiva Kleinerman, Ariel Rosenfeld, and Sarit Kraus. 2018. Providing Explanations for Recommendations in Reciprocal Environments. In *Proceedings of the 12th ACM Conference on Recommender Systems* (Vancouver, British Columbia, Canada) *(RecSys '18)*. ACM, New York, NY, USA, 22–30.

[31] Pigi Kouki et al. 2019. Personalized Explanations for Hybrid Recommender Systems. In *Proceedings of the 24th International Conference on IUI* (Marina del Ray, California) *(IUI '19)*. ACM, New York, NY, USA, 379–390.

[32] Todd Kulesza et al. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of the 20th International Conference on IUI* (Atlanta, Georgia, USA) *(IUI '15)*. ACM, New York, NY, USA, 126–137.

[33] Geneviève L. Lavigne, Robert J. Vallerand, and Paule Miquelon. 2007. A motivational model of persistence in science education: A self-determination theory approach. *European Journal of Psychology of Education* 22, 3 (01 Sep 2007), 351.

[34] Zachary C Lipton, David C Kale, Charles Elkan, and Randall Wetzell. 2015. Learning to diagnose with LSTM recurrent neural networks. *arXiv preprint arXiv:1511.03677* (2015).

[35] Yanjin Long and Vincent Aleven. 2017. Enhancing learning outcomes through self-regulated learning support with an Open Learner Model. *User Modeling and User-Adapted Interaction* 27, 1 (01 Mar 2017), 55–88.

[36] Xin Ma. 1999. A Meta-Analysis of the Relationship between Anxiety toward Mathematics and Achievement in Mathematics. *Journal for Research in Mathematics Education* 30, 5 (1999), 520–540.

[37] Andrew Mabbott and Susan Bull. 2006. Student Preferences for Editing, Persuading, and Negotiating the Open Learner Model. In *Intelligent Tutoring Systems*. Springer Berlin Heidelberg, Berlin, Heidelberg, 481–490.

[38] Millecamp Martijn, Cristina Conati, and Katrien Verbert. 2022. "Knowing Me, Knowing You": Personalized Explanations for a Music Recommender System. *User Modeling and User-Adapted Interaction* 32, 1–2 (apr 2022), 215–252.

[39] Bruce M McLaren and Seiji Isotani. 2011. When is it best to learn with all worked examples?. In *International Conference on AIED*. Springer, 222–229.

[40] Bruce M McLaren, Sung-Joo Lim, and Kenneth R Koedinger. 2008. When and how often should worked examples be given to students? New results and a summary of the current state of research. In *Proceedings of the 30th annual conference of the cognitive science society.* 2176–2181.

[41] Martijn Millecamp et al. 2019. To Explain or Not to Explain: The Effects of Personal Characteristics When Explaining Music Recommendations. In *24th International Conference Proceedings on IUI* (Marina del Ray, California) *(IUI '19)*. ACM, New York, NY, USA, 397–407.

[42] Martijn Millecamp et al. 2020. What's in a User? Towards Personalising Transparency for Music Recommender Interfaces. In *Proceedings of the 28th ACM Conference on UMAP* (Genoa, Italy). ACM, New York, NY, USA, 173–182.

[43] Martijn Millecamp, Cristina Conati, and Katrien Verbert. 2021-07-09. Classifeye: classification of personal characteristics based on eye tracking data in a recommender system interface.

[44] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38.

[45] Sidra Naveed, Tim Donkers, and Jürgen Ziegler. 2018. Argumentation-Based Explanations in Recommender Systems: Conceptual Framework and Empirical Results. In *26th Conference on UMAP* (Singapore, Singapore). ACM, New York, NY, USA, 293–298.

[46] James A. Overton. 2011. In *Scientific Explanation and Computation.* 41–50.

[47] Yasemin Abali Öztürk and Çavuş Şahin. 2014. The Effects of Alternative Assessment and Evaluation Methods on Academic Achievement, Persistence of Learning, Self-Efficacy Perception and Attitudes.

[48] Grant A Pignatiello, Richard J Martin, and Ronald L Hickman Jr. 2020. Decision fatigue: A conceptual analysis. *Journal of Health Psychology* 25, 1 (2020), 123–135.

[49] Kaśka Porayska-Pomsta et al. 2018. Adolescents' Self-regulation During Job Interviews Through an AI Coaching Environment. In *AIED*. Springer International Publishing, Cham, 281–285.

[50] Vanessa Putnam and Cristina Conati. 2019. Exploring the Need for Explainable Artificial Intelligence (XAI) in Intelligent Tutoring Systems (ITS). In *IUI Workshops*.

[51] Johnmarshall Reeve, Hyungshim Jang, Pat Hardre, and Mafumi Omura. 2002. Providing a rationale in an autonomy-supportive way as a strategy to motivate others during an uninteresting activity. *Motivation and emotion* 26, 3 (2002), 183–207.

[52] Sebastiaan Rothmann and Elize Coetzer. 2003. The Big Five Personality Dimensions and Job Performance. *South African Journal of Industrial Psychology* 29 (10 2003), 68–74. https://doi.org/10.4102/sajip.v29i1.88

[53] Markel Sanz Ausin et al. 2020. Exploring the Impact of Simple Explanations and Agency on Batch Deep Reinforcement Learning Induced Pedagogical Policies. In *AIED*. Springer International Publishing, Cham, 472–485.

[54] James Schaffer et al. 2019. I Can Do Better than Your AI: Expertise and Explanations. In *Proceedings of the 24th International Conference on IUI* (Marina del Ray, California). ACM, New York, NY, USA, 240–251.
[55] Nava Tintarev and Judith Masthoff. 2012. Evaluating the effectiveness of explanations for recommender systems. *User Modeling and User-Adapted Interaction* 22, 4 (01 Oct 2012), 399–439.
[56] Chun-Hua Tsai and Peter Brusilovsky. 2019. Designing Explanation Interfaces for Transparency and Beyond.. In *IUI Workshops*.
[57] Danding Wang et al. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk). ACM, New York, NY, USA, 1–15.
[58] Xi Yang et al. 2020. Student Subtyping via EM-Inverse Reinforcement Learning. In *Proceedings of the 13th International Conference on EDM 2020, Fully virtual conference, July 10-13, 2020.* International Educational Data Mining Society.