Saudi Journal of Humanities

المجلة السعودية للعلوم الإنسانية
Saudi Journal of Humanities

thirteenth edition

Volume (4) Issue (4)

(1-25) 2026

Review article: The Effectiveness of Explainable AI-Based Adaptive Simulation in Developing Sustainable Judgment and Trust Calibration in Algorithmic Decisions Within Changing Environmental Scenarios Among School Students

Mustafa Abdulkhader Jassim
(BABYLON EDUCATION DIRECTORATE/EDUCATIONAL RESEARCH AND STUDIES DIVISION)

Published on: Published online on 10, November 2026

Abstract :

This study investigated the impact of an explainable artificial intelligence (XAI) adaptive simulation on building sustainable judgment and trust calibration in algorithmic decisions in varying environmental scenarios with school students. The study used a two-group pretest–posttest design with 160 students (80 students in XAI-adaptive and 80 students in control group) from 8th to 11th grade. Three decision scenarios (stable conditions, drought shift and extreme weather) were administered to all students, and they were evaluated pre- and post-tests on sustainable judgment, trust calibration, overtrust, undertrust, AI literacy and explanation understanding. The XAI-adaptive group was the only one that was explained and adapted during the simulation. Paired and Welch t tests, Cohen's d, and analysis of covariance were used to analyze the data. The XAI-adaptive group showed a 18.29 point increase in sustainable judgment compared to the control group (d = 2.07), and a 21.65 point increase in trust calibration compared to the control group (d = 2.82). Overtrust decreased by 11.54 points compared to 2.33, and undertrust decreased by 9.68 points compared to 1.01. Improvements were also greater in the areas of AI literacy (20.35

vs 5.63) and explanation understanding (25.52 vs 4.69). The accuracy of the mean scenario was higher in the XAI-adaptive group (74.66) than in the control group (54.84), with the difference being the most pronounced in extreme weather (21.51 points). All differences were significant at $p < .001$ and were not due to pre test scores. The results suggest that explanation-supported adaptive simulation is effective to support students' appropriate judgment of algorithmic advice when conditions change, but the large effects warrant further replication with follow-up.

Keywords:

explainable AI, adaptive simulation, trust calibration, sustainable judgment, school students, environmental scenarios.

1. Introduction

AI is no longer just a part of the school curriculum, but a key element of the way students access information, get feedback, and make decisions. Adolescents are faced with many systems that recommend, predict and rank in their daily life and educational researchers talk about the opportunities and the dependencies that this shift entails. With the use of algorithmic tools becoming commonplace, students are expected to not only use these tools, but also to assess when the output of these tools is to be trusted. Therefore, the education of young people in algorithmic decision environments is an urgent educational issue (Chen et al., 2020).

This worry is more pronounced in the environmental sectors, where algorithms are being increasingly employed for prediction, resource management and adaptation to climate change. Recommendations in these areas are consequential and conditions that they describe can change rapidly, such as from a stable season to drought or a storm. These systems will be passed down to and then used by young people, and the environment is

changing, which means that young people need to practice assessing algorithmic advice in a changing environment (Rolnick et al., 2022).

In the field of sustainable development education, it is important to teach students the skills necessary to make informed choices and act responsibly in order to ensure environmental integrity. Sustainable judgment is the ability to make decisions about the long-term environmental and social impacts of choices, particularly when conditions change, as demonstrated in this study. This is a key component of sustainability education, and is now more often than not being used in tandem with algorithmic recommendations than without them (UNESCO, 2020).

One challenge that continues to be a hurdle is the lack of transparency in many AI systems. Explainable AI is a key step towards responsible AI and seeks to make the behaviour of AI models understandable to human users. However, the value of explanation is only useful when people can use it to make decisions on when to follow and when to question a system, a matter of human learning and not just of model design (Arrieta et al., 2020).

Although the use of XAI and educational AI has been expanding, there is limited knowledge regarding whether explanation-supported, adaptive learning environments can be used to support the calibration of school students' trust in algorithmic advice in varying environmental contexts. This study thus examines the effectiveness of an XAI-based adaptive simulation on students' sustainable judgement, trust calibration (overtrust and undertrust), AI literacy, understanding of the explanation, and accuracy in decision-making in the three scenarios: stable, drought, and extreme weather. The paper is followed by a literature review and hypotheses, methodology, results, discussion and conclusion.

2. Literature Review

2.1 Explainable AI and Interpretability

Explainable AI research has led to a large family of methods that provide insights into the processes that lead to model predictions, including inherently interpretable models as well as post hoc methods like feature attribution and example-based explanations. This literature review highlights the differences in the scope, fidelity, and the type of insight gained from the different methods, making it difficult to find a method that works for all audiences. For learners, this involves choosing explanations and presenting them in a way that they can interpret in a realistic way (Linardatos et al., 2021).

Technical measures are not the only ones needed to evaluate an explanation. The literature suggests that the goodness of the explanation, user satisfaction, quality of the user's mental model, curiosity, trust, and joint human–AI performance are all measures of explanation goodness. This more general perspective sees explanations as agents of change to what people understand and do, and directly motivates the measurement of understanding of explanations and the outcomes of trust in an educational context (Hoffman et al., 2023).

2.2 Explainable AI in Education

In the field of education, explain ability is particularly important as decisions made by AI systems for and about learners have a significant impact on their developing minds and must be comprehensible to both teacher and learner. In the field of education, the literature on explainable AI claims that explanations should be tailored to the needs of the education stakeholders and not just borrowed from other fields. It also highlights unanswered questions regarding the role of explanations in classrooms in terms of learning and decision making (Khosravi et al., 2022).

This is supported by ethical guidelines for using AI in education that prioritize transparency, accountability, and respecting students' autonomy. If students don't understand the rationale behind a system's recommendations, they are not in a position to challenge it, and explanation can be seen as a way of respecting students as active agents.

Creating explanation as an educational tool is thus a pedagogical and an ethical decision (Holmes et al., 2022).

2.3 Trust Calibration and Appropriate Reliance

Trust calibration is the matching of trust in an AI system with the reliability of the system. Overtrust is when people take in bad advice, and undertrust is when they don't take in good advice. Experimental results demonstrate that the accuracy and trust of a model can be affected by displaying a model's confidence (and to a lesser degree, explanations), while explanations do not necessarily lead to better calibration (Zhang et al., 2020).

It is important to the type of explanation. Research on clinical decision support reveals that the impact of the various explanation classes on trust calibration varies and some explanation classes may lead to an increase in trust without an increase in the appropriateness of the trust. Based on this evidence, it is suggested that the educational designs should be geared towards appropriate reliance and not trust per se (Naiseh et al., 2023).

Appropriate reliance has been conceptualized as relying on AI advice when it is right, but not relying on it when it is wrong – a distinction between blind acceptance and discerning use. Studies of this idea suggest that explanations can influence reliance, but that this influence varies depending on how people interact with explanations. This framing is reflected in the separate measure of overtrust and undertrust in the present study (Schemmer et al., 2023).

2.4 Adaptive Learning Systems and AI in Education

Adaptive learning systems customize content, feedback, or difficulty to the learner and a systematic review of the literature reveals that this area has seen a lot of growth and that there is mixed evidence regarding learning outcomes. The mapping identifies the need for

research to link adaptive features to cognitive and behavioral outcomes, and not just engagement. Thus, Adaptive adjustment is an important design feature of the current intervention (Kabudi et al., 2021).

A distinction has been made between three paradigms of AI in education: AI-directed, AI-supported and AI-empowered. They vary in the part played by the learner, from one of receiving instruction to one who is a partner and shares control with the system. An intervention that requires students to examine explanations and modify a simulation is at the end of the spectrum where the learner is empowered (Ouyang & Jiao, 2021).

2.5 AI Literacy

AI literacy is a collection of skills that allow people to critically assess AI technologies, interact and interact with AI, and apply AI as a tool. The framework encompasses an understanding of how AI functions, its constraints and limitations, and reasoning about its output, the skills required to tune the trust. Thus, it is only natural that AI literacy would be a natural companion outcome to trust calibration (Long & Magerko, 2020).

2.6 Research Gap and Hypotheses

The literature provides good conceptual support for explanation, adaptivity and literacy, and very little empirical research has placed school students in dynamically changing environmental situations, and assessed sustainable judgment and calibrated trust simultaneously. This gap led to four hypotheses being tested. H1: The XAI-adaptive group will make more gains in sustainable judgment than the control group. H2: it will be more successful in terms of trust calibration and have larger decreases in overtrust and undertrust. H3: it will be more accurate in decision making in different scenarios including extreme weather. H4: It will improve its AI literacy and understanding of explanations more.

3. Methodology

3.1 Research Design

This study adopted a two-group pretest–posttest design, with one group being the simulation group with XAI and the other group being the control group. Pretest and posttest scores and the three decision scenarios were identical for both groups. The explanation layer and adaptive changes were only given to the XAI-adaptive group in the simulation. Sustainable judgment, trust calibration, overtrust, undertrust, AI literacy, understanding of the explanations, and accuracy of the decisions made in the scenarios were the dependent variables.

3.2 Participants

A total of 160 school students (80 in each grade) were used in the sample. The mean ages were 14.70 years (SD = 1.39) for the XAI-adaptive group and 14.96 years (SD = 1.40) for the control group, with 89 students being female and 71 students being male. The secondary school students were selected as they are already regular users of algorithmic tools and AI literacy education in K-12 schools is still in development (Casal-Otero et al., 2023).

3.3 The XAI Adaptive Simulation

The intervention was a simulation where students were asked to make decisions based on algorithmic recommendations in three scenarios of the environment: stable, drought shift and extreme weather. The XAI-adaptive group received a recommendation with an

explanation and the simulation made the changes based on the student's choices. The system saved the simulation time, the number of explanation views and the number of adaptive adjustments. The design is based on the perspective that in the field of education, AI should serve the students and also conduct research on the challenges and conditions of AI use (Hwang et al., 2020).

3.4 Measures

Sustainable judgment, trust calibration, AI literacy and explanation understanding were measured on a 0–100 scale with higher scores signifying better performance at pre-test and post-test. Overtrust and undertrust were also noted at both time points and lower scores represent better calibration. For each scenario, the accuracy of the decision, the time it took for the student to make the decision (in seconds), and whether or not the student used an explanation and whether or not the student changed her decision after reading an explanation were noted. Self-reported confidence and a trust appropriateness score was also gathered. The AI literacy and explanation understanding measures are based on the principle of the AI literacy course evaluation, which evaluates conceptual understanding in addition to its application (Kong et al., 2021).

3.5 Procedure and Data Analysis

The pre-test was administered to students before the three scenarios, and then the post-test was administered after the three scenarios. Gain scores were calculated by subtracting the pre test scores from the post test scores. Independent-samples t tests and chi-square

tests were used to check for baseline equivalence. Paired samples t tests were used to test within group change, between group differences in gains were tested with Welch t tests and Cohen's d. Analysis of covariance (ANCOVA) with the pre-test as covariate was also conducted for each outcome since there were differences in baseline scores. The differences between the scenarios were examined using Welch t tests and relationships between engagement indicators and gains were examined using Pearson correlations. Significance level was .05 and the entire data set in Excel (160 participants and 480 scenario trials) was used for all analyses.

4. Results

4.1 Baseline Equivalence

The participant characteristics are summarized in Table 1. There were no significant differences between the groups for age, grade distribution, gender or prior exposure to AI, but there was a trend ($p = .070$) for more Grade 8 students in the XAI-adaptive group. The number of students who had not been exposed to prior AI was 38 (23.8%) and this was considered as a separate category.

Table 1. Participants' characteristics according to group

Characteristic	XAI-adaptive (n = 80)	Control (n = 80)	Test
Age, M (SD)	14.70 (1.39)	14.96 (1.40)	$t(158) = -1.19$, $p = .236$
Grade 8, n (%)	37 (46.3)	26 (32.5)	$\chi^2(3) = 7.05$, $p = .070$

Grade 9, n (%)	19 (23.8)	18 (22.5)	
Grade 10, n (%)	13 (16.3)	27 (33.8)	
Grade 11, n (%)	11 (13.8)	9 (11.3)	
Female, n (%)	42 (52.5)	47 (58.8)	$\chi^2(1) = 0.41$, $p = .524$
Male, n (%)	38 (47.5)	33 (41.3)	
Prior AI exposure: High, n (%)	13 (16.3)	12 (15.0)	$\chi^2(3) = 1.07$, $p = .784$
Moderate, n (%)	26 (32.5)	21 (26.3)	
Low, n (%)	24 (30.0)	26 (32.5)	
Missing, n (%)	17 (21.3)	21 (26.3)	

As shown in Table 1, the two groups were similar with respect to the demographic variables, thus allowing the comparison of their results. There was a small bias in the grade distribution that should be taken with a pinch of salt, but was not significant. Next, pre-test scores were examined to see if there was complete equivalence on each outcome.

4.2 Pre-Post Change Within Each Group

The pre-test and post-test results are given in Table 2. Both groups saw improvements on all outcomes and all within-group differences were significant at $p < .001$, with the exception of the control group's decrease in undertrust, which was significant at $p = .016$. The enhancements of the XAI-adaptive group were much more significant in all instances.

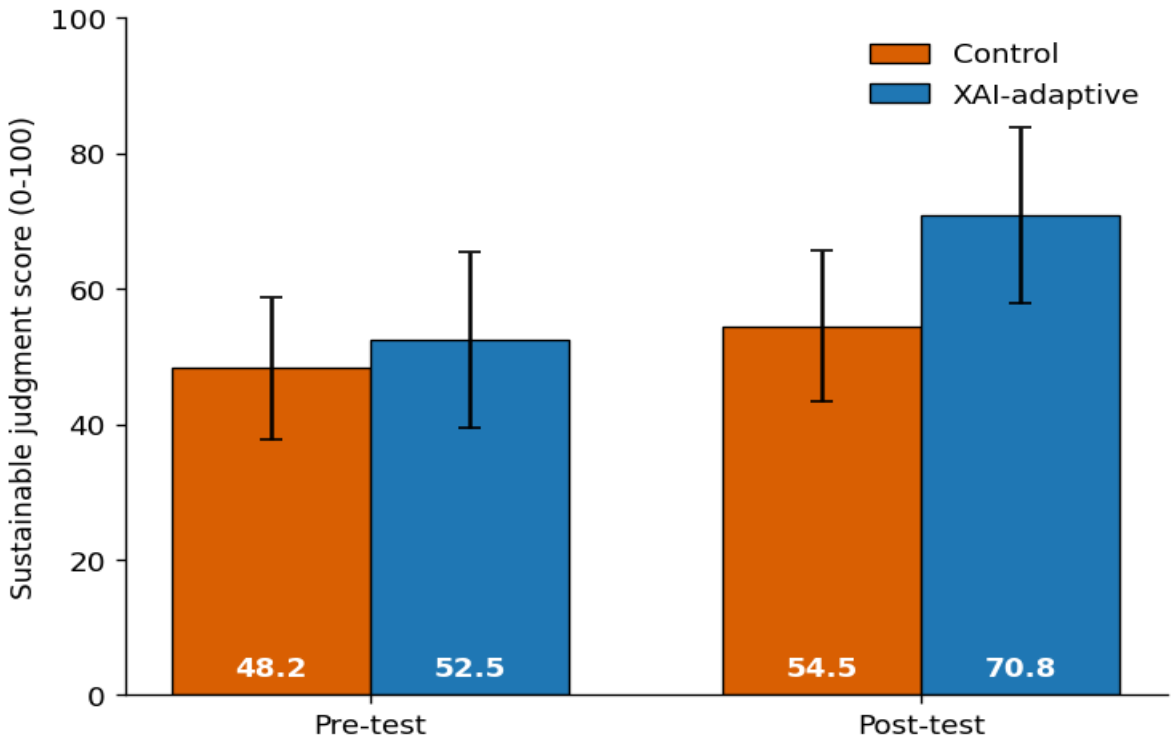
Table 2. Pre-test and post test scores for each group

Outcome	**Group**	**Pre M (SD)**	**Post M (SD)**	**Change**	**t(79)**	**d_z**
Sustainable judgment	XAI	52.46 (13.06)	70.75 (12.99)	+18.29	24.25	2.71
	Control	48.24 (10.46)	54.47 (11.09)	+6.23	11.74	1.31
Trust calibration	XAI	45.66 (13.44)	67.30 (15.09)	+21.65	29.11	3.25

	Control	49.38 (14.19)	53.88 (14.80)	+4.50	7.38	0.83
Overtrust	XAI	36.28 (13.53)	24.74 (13.56)	−11.54	−23.34	−2.61
	Control	39.87 (14.31)	37.54 (14.93)	−2.33	−5.10	−0.57
Undertrust	XAI	31.29 (10.62)	21.62 (11.10)	−9.68	−19.39	−2.17
	Control	30.63 (12.50)	29.62 (12.37)	−1.01	−2.46*	−0.28
AI literacy	XAI	47.16 (12.29)	67.51 (13.38)	+20.35	32.13	3.59
	Control	45.67 (14.03)	51.29 (15.10)	+5.63	9.62	1.08
Explanation understanding	XAI	40.89 (14.36)	66.41 (14.33)	+25.52	35.62	3.98
	Control	39.01 (13.22)	43.70 (13.92)	+4.69	8.04	0.90

Note. All $p < .001$ except *$p = .016$.

Table 2 indicates that the XAI-adaptive group experienced moderate to significantly higher scores for all outcomes. For instance, sustainable judgment increased from 52.46 to 70.75 and trust calibration went up from 45.66 to 67.30. The smaller gains of the control group are in line with practice and repeated testing effects and serve as the standard for evaluating the effectiveness of the intervention. The groups also differed at pre-test on sustainable judgment (52.46 vs 48.24, $p = .025$) which is discussed in the next section.



Figure

1. Sustainable judgment before & after (bar chart)

The sustainable judgement is shown graphically in figure 1. The bars indicate that the two groups began close together and diverged at post-test, with the XAI-adaptive group having an average score of ~71 and the control group an average score of ~54. The standard deviations (error bars) suggest that the distributions of the posttests overlap only slightly. The figure thus represents the same pattern as the Table 2, in a manner which is immediately apparent.

4.3 Between-Group Comparison of Gains

The gain scores are directly compared in Table 3. The XAI-adaptive group improved significantly more than the control group on all six outcomes (all $p < .001$) with the ANCOVA adjusted differences being nearly the same as the raw differences. Hypotheses H1, H2 and H4 were thus confirmed.

Table 3. The difference between the mean gain scores for the two groups.

Outcome	XAI gain M (SD)	Control gain M (SD)	Difference [95% CI]	Welch t	Cohen's d	ANCOVA adj. diff.
Sustainable judgment	18.29 (6.75)	6.23 (4.75)	12.06 [10.23, 13.88]	13.07	2.07	12.48
Trust calibration	21.65 (6.65)	4.50 (5.45)	17.15 [15.25, 19.05]	17.84	2.82	17.10
Overtrust	−11.54 (4.42)	−2.33 (4.09)	−9.21 [−10.54, −7.88]	−13.67	−2.16	−9.29
Undertrust	−9.68 (4.46)	−1.01 (3.67)	−8.67 [−9.94, −7.39]	−13.42	−2.12	−8.63
AI literacy	20.35 (5.66)	5.63 (5.23)	14.72 [13.02, 16.43]	17.08	2.70	14.72
Explanation understanding	25.52 (6.41)	4.69 (5.22)	20.83 [19.00, 22.66]	22.54	3.56	20.95

Note. All $p < .001$. Negative values for overtrust and undertrust indicate improvement. Adjusted differences come from ANCOVA with the pre-test as covariate.

As can be seen in Table 3, the biggest advantage for the intervention was for explanation understanding (20.83 points) and for trust calibration (17.15 points), which are the most direct outcomes related to the explanation layer. Both overtrust and undertrust were decreased in the XAI-adaptive group, this resulted in students being less likely to accept

bad advice and less likely to reject good advice. The agreement between raw and adjusted differences indicates that there is no agreement between them when it comes to the baseline difference in sustainable judgment.

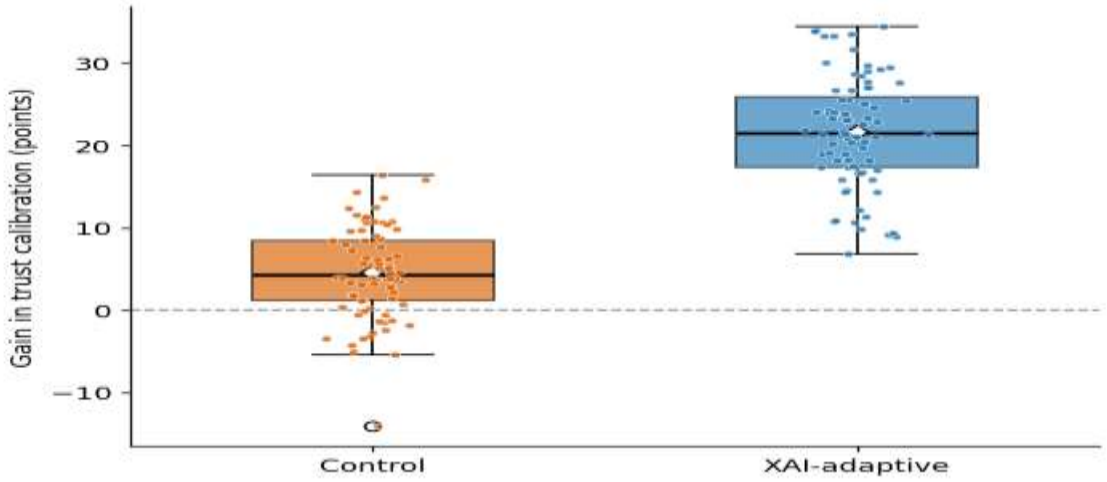


Figure 2. The graph shows a box plot of the results for trust- calibration by group.

The distribution of the trust- calibration gains is shown in Figure 2. The median gain of the XAI-adaptive group was 21.55 points (interquartile range 17.40 to 25.88) and all students gained points ranging between 6.8 and 34.4. The median for the control group was a gain of 4.25 points (IQR 1.20 to 8.50) with individual gains ranging from −14.1 to 16.4 and some students making a loss. The two distributions hardly overlap, and the difference between the two groups is not due to a few extreme cases, but rather to a shift in the distribution.

4.4 Performance Across Environmental Scenarios

The accuracy of the decisions made and the behaviour in each scenario is reported in Table 4. The XAI-adaptive group did better than the control group in all three scenarios (all $p < .001$), thus supporting H3. The accuracy drop was more pronounced for the extreme weather conditions in both groups, but was less with XAI support (8.64 points) compared to without XAI support (11.00 points).

Table 4. The accuracy of the scenarios and decision behaviour of the groups.

Scenario	Group	Accuracy M (SD)	Decision time, s M (SD)	Used explanation (%)	Changed decision (%)	Accuracy test
Stable	XAI	79.35 (10.23)	47.56 (17.21)	76	51	t = 11.01, d = 1.74
	Control	60.20 (11.71)	32.66 (10.01)	0	0	
Drought shift	XAI	73.92 (10.98)	45.74 (16.31)	76	40	t = 11.00, d = 1.74
	Control	55.14 (10.61)	30.99 (10.37)	0	0	
Extreme weather	XAI	70.71 (13.07)	45.78 (15.69)	74	31	t = 10.23, d = 1.62
	Control	49.20 (13.54)	30.34 (10.16)	0	0	
Overall	XAI	74.66 (6.85)†	46.36 (16.37)	75.4	40.8	t = 18.78, d = 2.97
	Control	54.84 (6.48)†	31.33 (10.18)	0	0	

Note. All $p < .001$. †Participant-level mean across the three scenarios. The control group had no explanation layer by design.

As can be seen in Table 4, the XAI advantage was 19.15 points in the stable scenario, 18.78 points in the drought shift and 21.51 points in extreme weather, meaning the advantage did not decrease when the weather shifted. The time to decision (46.36 vs. 31.33 seconds) indicates that the students in XAI took a longer time to process each decision, suggesting more deliberate processing, while the students in XAI used explanations in

approximately three-quarters of the trials. Within XAI trials, accuracy was higher when an explanation was used (75.49 vs 72.12, Welch t = 2.14, p = .034). This comparison at the trial level is not intended to be read as a comparison of the trials, and should be used with caution.

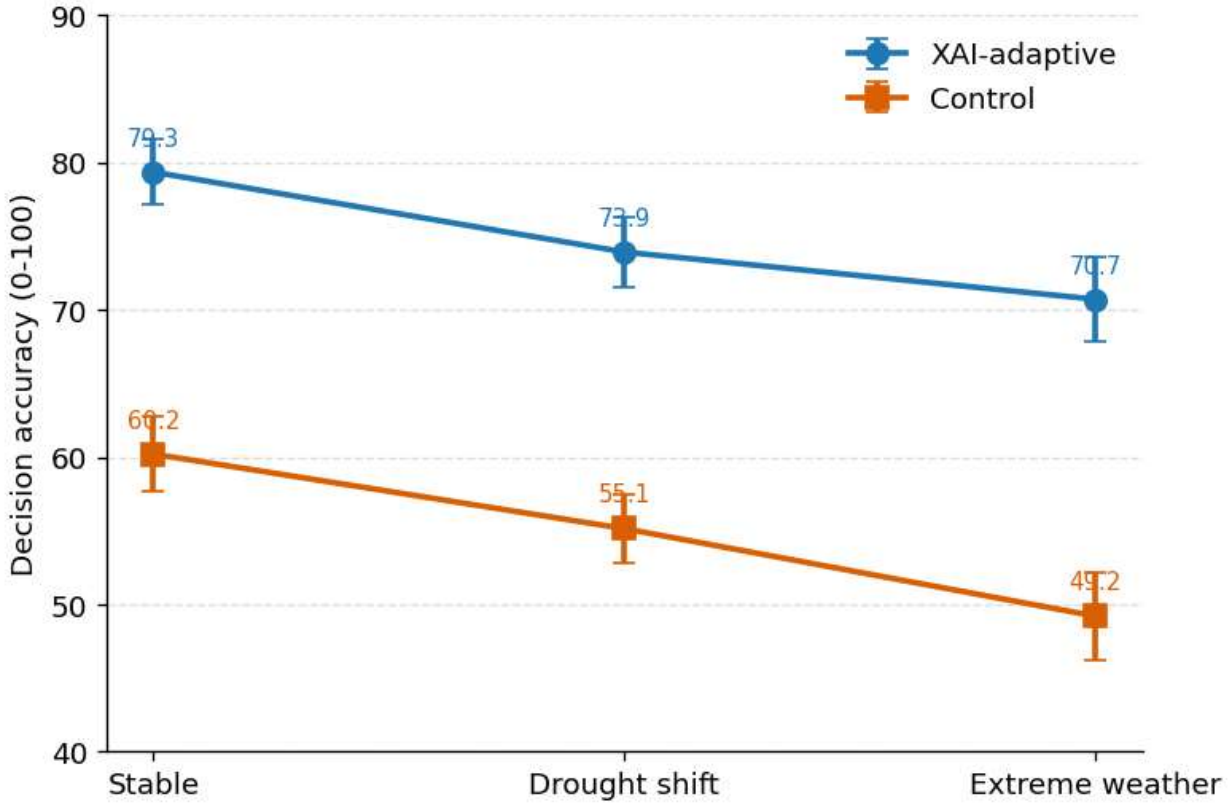


Figure 3. The accuracy of the answers for each scenario (line chart)

The means of the scenarios are plotted in Figure 3 with 95% confidence intervals. The two lines are approximately parallel and both are decreasing from the stable to the extreme weather scenario and the difference between the two lines is almost constant. The confidence intervals do not overlap in any case, which substantiates the advantage is solid in both cases. The unsupported students were more impacted by environmental change as revealed by the slightly steeper fall in the control line.

4.5 Individual Change and Engagement

Post-test self-reported confidence was higher in the XAI-adaptive group (72.83, SD = 10.44) than in the control group (57.38, SD = 13.16; $t = 8.23$, $p < .001$, $d = 1.30$). The trust appropriateness score was also higher (50.93, SD = 17.85 vs 30.07, SD = 16.00; $t = 7.78$, $p < .001$, $d = 1.23$), which indicates that the confidence gain was accompanied by more appropriate trust.

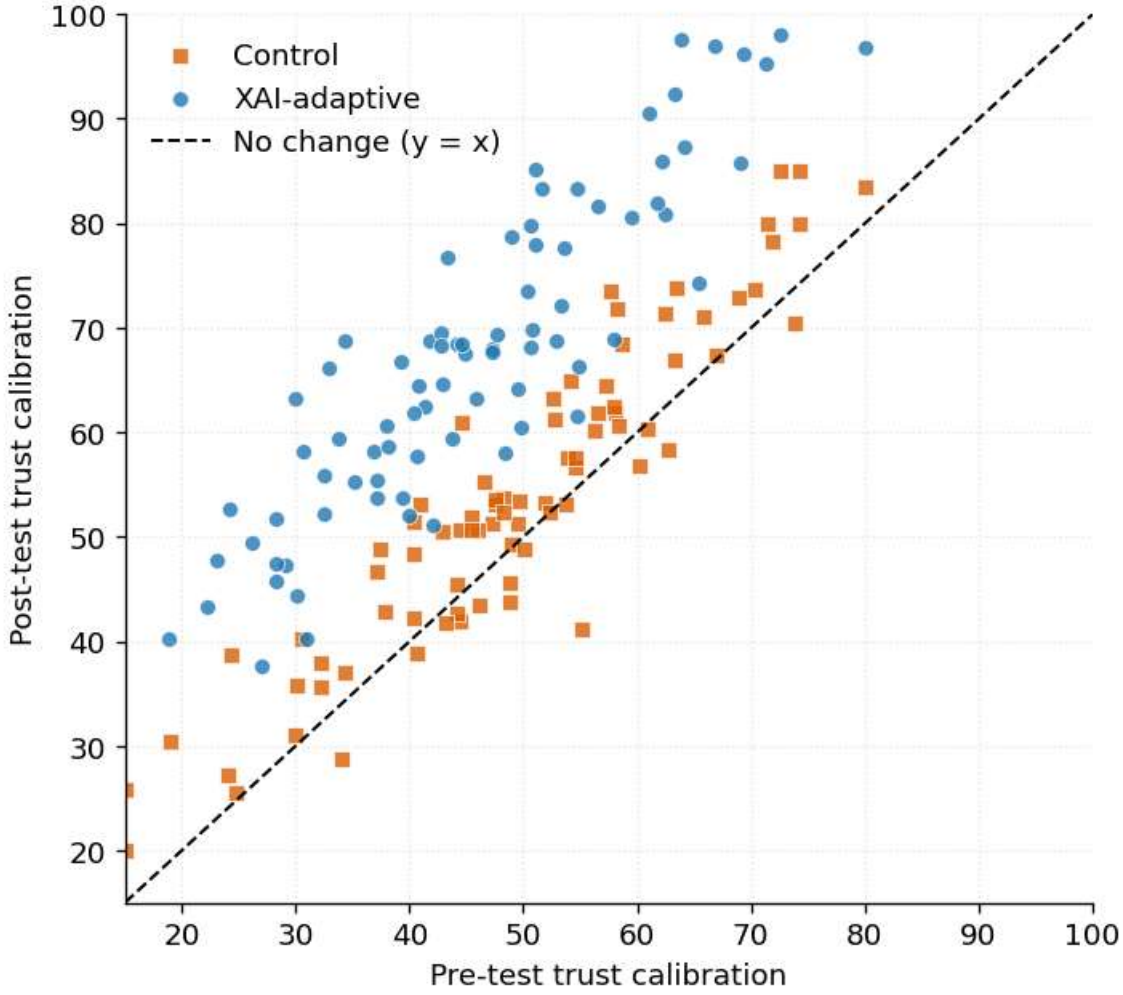


Figure 4. The pre-test vs post-test trust calibration (scatter plot).

Figure 4 shows each student's pre-test and posttest trust calibrations plotted on a line of no change. All 80 XAI-adaptive students (100%) are above the diagonal, while 64 out of 80 control students (80%) are above the diagonal and the XAI cluster is much further above the diagonal. The pre-test and post-test scores were highly correlated for both groups ($r = .90$ and $r = .93$) and the XAI group as a whole moved up in relative standing. The figure demonstrates that the benefit was widespread and did not depend on the students starting with a low level of trust calibration.

In the XAI-adaptive group, students on average spent 38.80 minutes ($SD = 8.30$) in the simulation and 11.10 minutes ($SD = 3.79$) on explanations and 6.46 minutes ($SD = 2.96$) on adaptive adjustments. The explanation views and gains in trust calibration were not significantly related to gains, nor were adaptive adjustments and gains in trust calibration, nor were simulation time and gains in sustainable judgment ($r = -.13$, $p = .242$; $r = -.21$, $p = .062$; and $r = .06$, $p = .615$, respectively). There was no evidence of any differences in the amount of engagement between students explaining differences in their gains.

5. Discussion

The results of this study provide support for all four hypotheses. The students in the XAI-adaptive simulation improved significantly in sustainable judgment, trust calibration, AI literacy, and understanding explanations, decreased overtrust and undertrust, and increased their accuracy in all scenarios compared to the control group. The effects were substantial and remained after controlling for the pretest scores.

The enhancement of sustainable judgment and accuracy of scenarios aligns with the narrative of AI in education as described in reviews, which highlight the swift growth of AI-driven learning and emphasize the importance of aligning AI systems with measurable outcomes. The present study meets that call with the measurement of judgment and accuracy under varying conditions, and indicates that a combination of explanation and

adaptivity can help ensure the quality of the decisions made under varying conditions (Zhai et al., 2021).

The finding of reduced overtrust aligns with experimental results showing that providing explanations can lower overtrust in AI systems when making decisions. Students who were able to understand the reasons behind a recommendation were less likely to accept it uncritically, and the fact that the undertrust was also reduced indicates that providing an explanation did not just lead to greater scepticism. It assisted them to distinguish between giving and questioning advice (Vasconcelos et al., 2023).

Decision times in the XAI-adaptive group are consistent with work that has demonstrated that designs that promote analytical engagement with AI advice lead to less overreliance, thereby supporting the notion that the XAI-adaptive group is more likely to consider the AI's advice analytically. Explanation views and adaptive adjustments might have served as prompts to think before accepting a recommendation in this study but the lack of a dose–response relationship with engagement indicates that the structure of the task was more important than the level of engagement (Buçinca et al., 2021).

The accuracy results also have implications for the support of good joint performance in the explanations. Previous studies revealed that explanations do not always lead to complementary human–AI performance, and the small benefits of using explanations in this trial lend credence to this caution. The advantage at the group level was large, indicating that it was not just the explanations, but the entire adaptive design, where students were required to take action with the explanations, that led to the advantage (Bansal et al., 2021).

Improvements in AI literacy and understanding of explanations suggest that students gained knowledge that they could apply to AI systems in general and their limitations. This is important because literacy frameworks include critical evaluation and understanding of AI as a key skill and it is these skills that enable calibrated trust.

Calibration and literacy thus seem to be reinforcing each other as results of the intervention (Ng et al., 2021).

The results indicate that, in the classroom, students can learn to critically engage with algorithmic advice by means of simulations rather than just warnings. This is particularly significant as generative AI tools are introduced into schools, as they are likely to produce fluent outputs that can lead to uncritical acceptance, and teachers are encouraged to foster students' critical thinking and awareness of limitations along with the use of generative AI tools (Kasneci et al., 2023).

The impact of the effects is large, so it is necessary to be careful. The impacts beyond d = 2 are rare in educational research, and some critics suspect that the potential of AI for education is exaggerated and that the findings of these studies are not definitive. Before strong claims can be made about the durability of change, there is a need for replication with independent samples, longer follow-up and delayed post-tests (Selwyn, 2022).

There are a number of caveats. There was an imbalance in the groups in pre-test sustainable judgment and in grade distribution, but the results of the ANCOVA analysis were similar. Some of the difference may be due to the novelty and engagement of an interactive simulation and not just to explanation or adaptivity, as there was no comparable-time activity for the control group. Almost a quarter of the students had not been exposed to prior AI. There were no psychometric properties of the instruments in the data set and the outcomes were measured immediately after the intervention. Lastly, repeated trials were assumed to be independent in the analysis at the trial level.

Further studies are needed to distinguish between the effects of explanation and adaptivity by comparing simulation only, explanation only, and combined conditions. It should also examine the various types of explanations, as well as long-term retention. Research can be informed by the wider research programme on explainable AI, which has a strong focus on interdisciplinary research and the human aspects of explanation (Longo et al., 2024), to create explanations for young learners (Longo et al., 2024).

6. Conclusion

The purpose of this study was to determine if an explainable, adaptive simulation was effective in assisting school students to make better and more appropriate decisions when algorithmic advice was provided in a changing environment, and if students would trust the advice given by the algorithm. This aim is supported by the evidence of 160 students in two groups. The students who participated in the XAI-adaptive simulation made significant gains compared to the control group in sustainable judgment, trust calibration, AI literacy, and understanding the explanations. They also decreased overtrust and undertrust more, and this effect was observed even after taking into account initial levels of these traits.

The findings are complemented by the scenario results. In the stable, drought, and extreme-weather scenarios, the XAI-adaptive group made more accurate decisions and their benefit did not diminish as the scenarios became more challenging. Students in this group also took longer to make their decisions and about three-quarters of their decisions involved explanations, indicating that they were more thoughtful and careful about the algorithmic recommendations. Combined the results suggest that embedding explanation in an adaptive task can provide students with learning opportunities for when to trust and when to question an algorithm rather than increasing or decreasing their trust.

The conclusions should be taken with a grain of salt. The effects were unusually sizeable, the groups were pretested on sustainable judgment, the control condition did not have an alternative interactive activity, and outcome was only assessed immediately following the intervention. The design should be replicated with independent samples and the conditions of explanation only, adaptivity only and combination compared, and delayed post-tests should be used to test for retention in future studies. Explanation-supported adaptive simulations may still be an interesting approach for schools and curriculum designers to

use to prepare students for a future where algorithmic advice will be commonplace in the decisions they will be making in their environment and everyday life.

References

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

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Article 81. https://doi.org/10.1145/3411764.3445717

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1), Article 188. https://doi.org/10.1145/3449287

Casal-Otero, L., Catala, A., Fernández-Morante, C., Taboada, M., Cebreiro, B., & Barro, S. (2023). AI literacy in K-12: A systematic literature review. International Journal of STEM Education, 10, Article 29. https://doi.org/10.1186/s40594-023-00418-7

Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. IEEE Access, 8, 75264–75278. https://doi.org/10.1109/ACCESS.2020.2988510

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2023). Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. Frontiers in Computer Science, 5, Article 1096257. https://doi.org/10.3389/fcomp.2023.1096257

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Shum, S. B., Santos, O. C., Rodrigo, M. T., Cukurova, M., Bittencourt, I. I., & Koedinger, K. R. (2022). Ethics of AI in education: Towards a community-wide framework. International Journal of Artificial Intelligence in Education, 32, 504–526. https://doi.org/10.1007/s40593-021-00239-1

Hwang, G.-J., Xie, H., Wah, B. W., & Gašević, D. (2020). Vision, challenges, roles and research issues of artificial intelligence in education. Computers and Education: Artificial Intelligence, 1, Article 100001. https://doi.org/10.1016/j.caeai.2020.100001

Kabudi, T., Pappas, I., & Olsen, D. H. (2021). AI-enabled adaptive learning systems: A systematic mapping of the literature. Computers and Education: Artificial Intelligence, 2, Article 100017. https://doi.org/10.1016/j.caeai.2021.100017

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274

Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., & Gašević, D. (2022). Explainable artificial intelligence in education. Computers and Education: Artificial Intelligence, 3, Article 100074. https://doi.org/10.1016/j.caeai.2022.100074

Kong, S.-C., Cheung, W. M.-Y., & Zhang, G. (2021). Evaluation of an artificial intelligence literacy course for university students with diverse study backgrounds. Computers and Education: Artificial Intelligence, 2, Article 100026. https://doi.org/10.1016/j.caeai.2021.100026

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable AI: A review of machine learning interpretability methods. Entropy, 23(1), Article 18. https://doi.org/10.3390/e23010018

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1–16. https://doi.org/10.1145/3313831.3376727

Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Del Ser, J., Guidotti, R., Hayashi, Y., Herrera, F., Holzinger, A., Jiang, R., Khosravi, H., Lecue, F., Malgieri, G., Páez, A., Samek, W., Schneider, J., Speith, T., & Stumpf, S. (2024). Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. Information Fusion, 106, Article 102301. https://doi.org/10.1016/j.inffus.2024.102301

Naiseh, M., Al-Thani, D., Jiang, N., & Ali, R. (2023). How the different explanation classes impact trust calibration: The case of clinical decision support systems. International Journal of Human-Computer Studies, 169, Article 102941. https://doi.org/10.1016/j.ijhcs.2022.102941

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. Computers and Education: Artificial Intelligence, 2, Article 100041. https://doi.org/10.1016/j.caeai.2021.100041

Ouyang, F., & Jiao, P. (2021). Artificial intelligence in education: The three paradigms. Computers and Education: Artificial Intelligence, 2, Article 100020. https://doi.org/10.1016/j.caeai.2021.100020

Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., Luccioni, A. S., Maharaj, T., Sherwin, E. D., Mukkavilli, S. K., Kording, K. P., Gomes, C. P., Ng, A. Y., Hassabis, D., Platt, J. C., … Bengio, Y. (2022). Tackling climate change with machine learning. ACM Computing Surveys, 55(2), Article 42. https://doi.org/10.1145/3485128

Schemmer, M., Kuehl, N., Benz, C., Bartos, A., & Satzger, G. (2023). Appropriate reliance on AI advice: Conceptualization and the effect of explanations. Proceedings of the 28th International Conference on Intelligent User Interfaces, 410–422. https://doi.org/10.1145/3581641.3584066

Selwyn, N. (2022). The future of AI and education: Some cautionary notes. European Journal of Education, 57(4), 620–631. https://doi.org/10.1111/ejed.12532

UNESCO. (2020). Education for sustainable development: A roadmap. https://unesdoc.unesco.org/ark:/48223/pf0000374802

Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., & Krishna, R. (2023). Explanations can reduce overreliance on AI systems during decision-making. Proceedings of the ACM on Human-Computer Interaction, 7(CSCW1), Article 129. https://doi.org/10.1145/3579605

Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenic, A., Spector, M., Liu, J.-B., Yuan, J., & Li, Y. (2021). A review of artificial intelligence (AI) in education from 2010 to 2020. Complexity, 2021, Article 8812542. https://doi.org/10.1155/2021/8812542

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. Proceedings of the 2020

Conference on Fairness, Accountability, and Transparency, 295–305. https://doi.org/10.1145/3351095.3372852