ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

Mapping the Computable Boundaries of Instructional Activities in Digitalized Education: A Hybrid Human-AI Framework.

Jiayi Li[1,*]

[1] Academy for Applied Policy Studies and Education Futures, the Education University of Hong Kong. 999077, Hong Kong, China

SUMMARY: *Digitization of education can be used to record and process teaching activities continuously in platform logs, learning analyses, and generative models, but there is still a lack of specific thresholds that determine whether a teaching activity is suitable for computation or whether it needs teacher leadership. The 312 identified teaching activities in this paper are divided into activity units; six dimensions of data observability, rule dominance, feedback immediacy, situational complexity, relationship sensitivity and ethical creation load are constructed, and a computable index and hybrid human-machine judgment framework are proposed. Based on the above results, content presentation and task practice fall within the fully computable range; formative evaluation is mostly auxiliary computable; collaborative learning is relatively conditional computable; and activities such as social-emotional support and ethical creation have shown obvious signs of being human-dominated. The weighted F1 of the hybrid framework for boundary recognition was 0.845, an increase of 10.0 percentage points over the regular baseline, and the residual risk dropped to 2.9%. The above research has offered an empirical basis for determining the limits of teaching activities and the integration model for teachers' assessments, risk thresholds, and control transfers. The two basic ideas of this system are limited automation and auditable control; thus, it is difficult to determine precisely where and when these two factors will be affected.*

KEYWORDS: *Digitalization of education; Teaching activities; Computable boundary; Hybrid human-machine framework; Teaching responsibility*

1 Introduction

After a class is conducted in the digital platform, the teaching objects will no longer be restricted to textbooks, classroom discussions, and homework results. The learning management system saves login and stay records, submission/review behaviour, the intelligent assessment system saves answers, error types, and feedback chains, and the generative model adds questions, explanations, examples and rewrites to the processable text. The above data can help schools and teachers optimize lesson preparation, diagnosis and assessment, but also raise the following questions: which parts of teaching activities can be entrusted to algorithms for stable processing, and which parts should only serve as auxiliary materials for teachers' judgment? Based on the previous studies, artificial intelligence applications in higher education and intelligent teaching systems have focused mainly on learning prediction, personalised recommendations, automated assessment and teaching management. Teachers' roles in the system design and results analysis stages have been relatively reduced recently. The new era of "learner-centred, data-driven and

*kzuhaaaa@163.com
https://doi.org/10.65102/is2026873
Published: 30 April, 2026

human-machine collaboration" also needs research that moves away from general talks about the availability of technology to specific boundaries of teaching activities.

Generative artificial intelligence has extended the scope of this problem. It can present questions, correct the answers provided by students, add relevant case studies, summarise the students' answers, and offer various feedback formats promptly. Related studies have shown that large language models can reduce some teaching preparation costs and provide learners with immediate support; however, the quality of their output is affected by prompts, training corpora and scenario constraints, and new uncertainties such as factual errors, inconsistencies in evaluation, and lack of learning responsibility attribution can easily arise [3]. Systematic studies in the field of education have shown that artificial intelligence applications are now used in various ways to produce content, analyze learning, conduct automatic evaluation, provide dialogue coaching, etc., and predict behaviour. However, the above applications often need clear data structures, stable task rules and reproducible evaluation criteria to be effective. When teaching activities that involve emotional regulation, value judgment, peer relationships and creative expression, the computing system can only provide some signals, suggestions or candidate explanations, and thus is unable to take full responsibility for teaching.

Define the 'computable range' more precisely. Molenaar's proposed hybrid human-machine learning technology suggests that the value of artificial intelligence in education is to promote the development of people's cognitive and regulatory abilities, and control should be jointly exercised by learners, teachers, and systems. Research on humanistic learning analysis has also indicated that if students and teachers are not involved in the system design and learning analysis, artificial intelligence may easily generate a "black box" suggestion; users lose control over the interpretation of data and risk identification and action selection. The above studies provide a foundation for this paper, but there is still a lack of research on the specific level of teaching activities; most studies focus on "what AI can do" or "what role teachers should play", and few have divided teaching activities into measurable units and used unified indicators to describe their computability, risk load, and boundaries of human-machine control.

There are deficiencies in both the foundation and higher education stages. According to the research on higher education, artificial intelligence has been applied to support courses, generate feedback, recommend learning paths, assist in academic writing, etc., but there are considerable differences in data conditions and teaching responsibilities for these various applications [7]. K-12 studies have shown that intelligent mentors, chatbots, educational games and robot applications generally perform better in knowledge practice, behaviour prediction and improvement of learning experience; however, stricter teacher intervention is required for ethical safety, student rights and emotional support [8]. The same "feedback" activity can be conducted at different levels; real-time feedback on multiple-choice questions is highly computable, open-ended writing feedback requires teacher review, and guiding student comfort and value after experiencing failure cannot be simplified into text generation tasks. If these divisions are not made first, educational digitalisation will be unable to distinguish between recordable items and automated ones.

There are also boundary problems in ethics and institutions. Ethical research on educational artificial intelligence proposes that fairness, transparency, privacy, responsibility, and the protection of children should be taken into account before applying such systems. The UNESCO Guidelines for Generative Artificial Intelligence also recommend that educational applications be people-centred in approach and set restrictions on data privacy, age-appropriateness, tool validation, and instructional design [10]. The above requirements indicate that, in the course of computerising teaching activities, the accuracy of algorithms, efficiency benefits, high-risk value judgments, protection of vulnerable students, and teachers' ability to assume explanatory responsibility based on on-site relationships all need to be considered

simultaneously. The research on computable scope should be based on the same set of tools and in a way that is conducive to teaching.

The structure of teaching activities at the institution also does not fall under the division of "discipline" or "platform type". The structure of the mathematics class is concept presentation, example deduction, individual exercises, peer discussion, error diagnosis and emotional support; a writing task includes material retrieval, outline generation, argument development, peer evaluation, teacher comments and value expression. The first few types of activities generally have clear inputs, explicit rules, and measurable outputs; on the other hand, the latter types of activities are based on classroom interactions, student experiences and teachers' sense of time. Only the objects that need to be defined should be realised through the combination of specific teaching activities in terms of data, rules, feedback and responsibility. Only by dividing activities into relatively small components can we explain why the same AI system works reliably in objective tests, yet still needs teacher-led discussions on organisation and ethics, as well as psychological support for students.

The "computability" referred to in this article is limited to the three conditions: teaching activities can be stably represented as data objects, verifiable outputs can be generated through rules, models, or thresholds, and the outputs will not increase ethical and relational risks in the process of teaching decisions. The two are technical feasibility and teacher-judgement interfaces.

Based on the above problems, this paper puts forward a hybrid human-machine framework for educational digitalisation scenarios around "computable boundaries of teaching activities", and applies this problem to the configuration of programmable activity units, boundary indicators and control rights. The teaching activities of this study are divided into programmable activity units, and the six dimensions of boundary indices—data observability, rule dominance, feedback immediacy, situational complexity, relationship sensitivity, and ethical creation load—are calculated; analysis is then organized according to the four intervals of fully computable, auxiliary computable, conditionally computable, and human dominated. The three main contributions of this paper are: first, to provide a variable caliber for the computable range of teaching activities and to turn the concept of "computability" from an abstract judgment into a measurable activity code; second, to construct a hybrid human-computer judgment framework and integrate teacher review, risk thresholds, and control transfers into the same model; third, to verify the distribution boundaries of different teaching activities through scenario-based samples and provide a foundation for the limited deployment of intelligent teaching systems, division of teaching responsibilities, and protection of teacher professional judgment.

2 Methods

2.1 Instructional Activity Unitization and Data Construction

To avoid directly equating platform functions with teaching activities, this paper first recodes the common content presentation, evaluation feedback, teaching support, collaborative learning, social emotional learning and other objects in discussions about educational AI, and supplements ethical creation activities to form six types of teaching activity units. The sample sources are publicly available course teaching design texts, common learning management system log fields, automatic assessment question type descriptions, and teacher feedback scenario descriptions; no real student identity information or personally identifiable information will be included in the data. Each activity unit contains only activity objectives, input materials, observable behaviour, evaluation rules, feedback methods, teacher-student relationship clues and potential risk explanations. A total of 312 activity units were eventually created, including

56 content presentations, 64 task exercises, 58 formative assessments, 46 collaborative learning sessions, 38 social and emotional support activities, and 50 ethical creation activities. This way is in line with the model of learning evidence, task difficulty and feedback path in personalised education research, but this paper focuses on whether teaching activities can enter a stable calculation process and does not aim to predict academic performance as its primary objective [11].

The display of the activity unit should meet the requirements of "recordable, interpretable and verifiable". Figure 1 is the three-dimensional object space with computable boundaries for teaching activities, as shown in Figure 1.

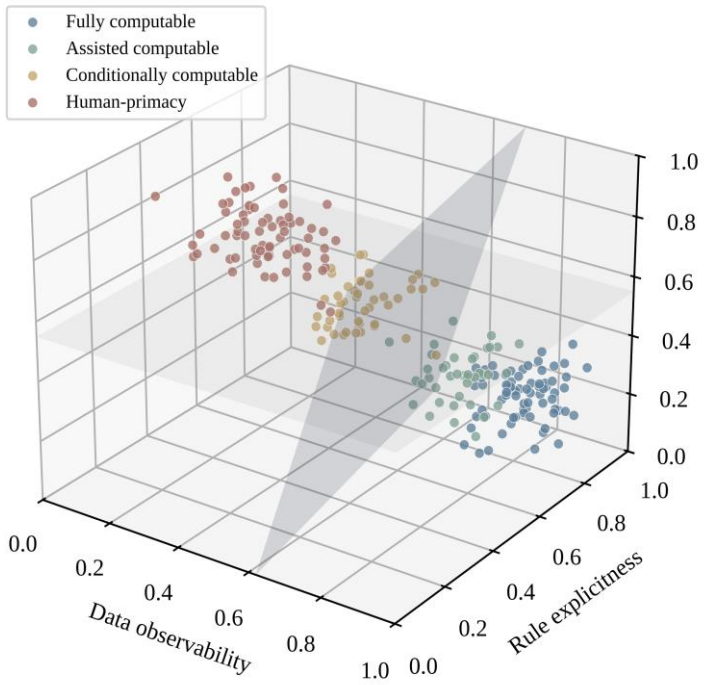

Figure 1: Three-dimensional Object Space with Computable Boundaries for Teaching Activities

As shown in Figure 1, the horizontal axis is data observability, the vertical axis is rule dominance, and the third vertical axis is human value load; activities in areas with high observability, high rule-basedness, and low value load are more likely to be automatically processed, while activities in areas with low observability, low rule-basedness, and high value load need to be managed by teachers. The above diagram shows the location of teaching activities in the three necessary conditions, reduces the obfuscation of complex activities by a single category, and enables boundary indices to monitor specific sources.

At the level of data organisation, the six variables in this paper are used to describe each activity unit: data observability, rule dominance, feedback immediacy, situational complexity, relationship sensitivity, and ethical creation load. The definitions of the variables are as follows: Table 1. The first three are the conditions that the algorithm can deal with, and the last three are teaching risk and human judgment load. The higher the observability of the data, the more complete the records of the activities on the platform will be; the more dominant the rules are, the more stable the task evaluation standards will be; and the higher the immediacy of feedback, the more likely it is that the system will generate useful feedback in a short period of time. The more complex the circumstances, the greater the sensitivity to interpersonal relationships, and the higher the ethical construction load, the more reliant an activity is on on-site observation,

trust in teacher-student relationships, value judgments, and thus the higher the risk associated with automated processing.

Table 1: Calculation of Boundary Variables and Measurement Scale for Teaching Activities

Symbol	Variable Name	Measurement Criteria	High-Value Meaning
d_i	Data observability	Whether platform logs, response records, textual evidence, and behavioral events are complete	The activity can be stably recorded
r_i	Rule explicitness	Whether scoring criteria, task constraints, and judgment rules are clear	The output can be reviewed
t_i	Feedback immediacy	Whether the system can generate usable feedback within a short time	The feedback chain is short
k_i	Contextual complexity	Whether the activity depends on classroom situations, disciplinary context, and individual background	On-site interpretation is required
s_i	Relational sensitivity	Whether the activity involves teacher–student trust, peer relationships, and emotional states	Teacher intervention is required
e_i	Ethical-creative load	Whether the activity involves value judgment, original expression, and responsibility attribution	Automated decision-making is not appropriate

The four anchor points for specific rating setting are as follows: 0-0.25 indicates that the activity is almost impossible to be recorded or regularized by the platform; 0.26-0.50 indicates that local records exist but require an additional explanation from the teacher; 0.51-0.75 indicates that the data and rules are generally applicable but still need to be reviewed; and 0.76-1.00 indicates that the records are complete, the rules are clear, and the feedback path is stable. To reduce the subjective experience drift of ratings, the coding personnel conducted a trial of 30 samples before formal labelling and concentrated the disputed samples in the "adjacent boundary" area for discussion. For example, information retrieval in inquiry-based discussions can achieve a high degree of data observability; the sensitivity of mediating conflicting viewpoints is included separately; grammar correction in writing feedback can be in a relatively high computable range, but argumentative value, creative expression, and student self-efficacy are included in ethical creative load or relationship sensitivity. Encoding can reduce the problem of complex behaviour being overlooked by the average.

Figure 2 shows the way of data organisation and sample construction, as shown in Figure 2.

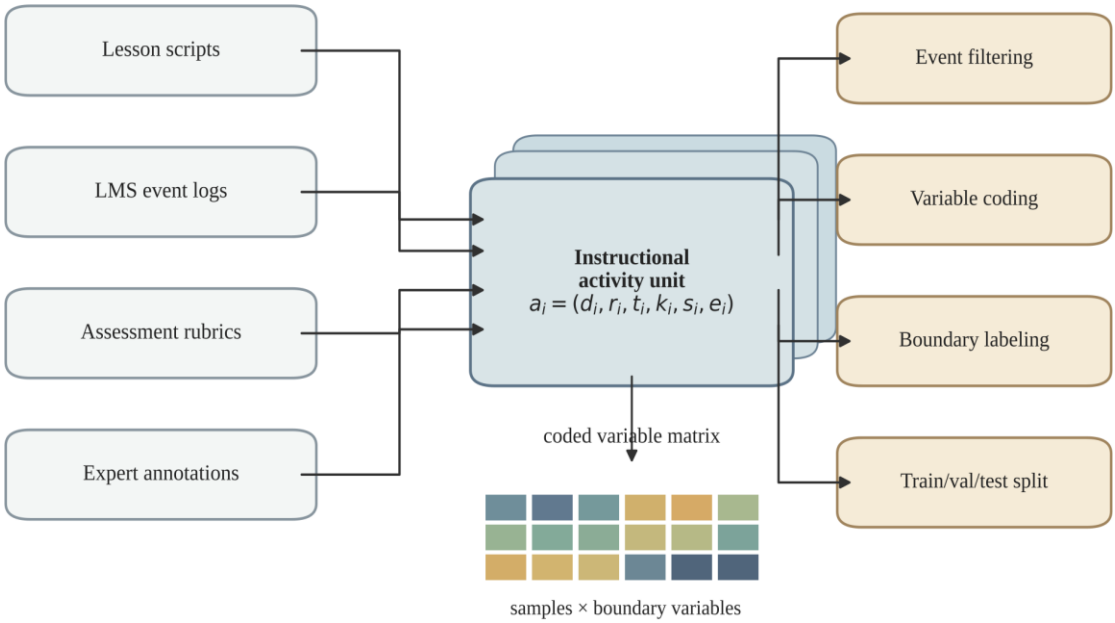

Figure 2: Data Organization and Activity Unit Encoding Mechanism

As shown in Figure 2, the publicly available instructional design text, learning platform log fields, evaluation rules and expert annotations are all collectively imported into the activity unit library; after event filtering, variable encoding, boundary annotation and sample partitioning, each activity unit is converted into a structured object that can be used in the model calculation. The above mechanism does not specify which data source or variable the model's output is derived from, and thus is difficult to identify the particular teaching evidence that supports the results.

The two encodings are as follows. The first round was independently scored by two researchers in the field of educational technology according to Table 1, and the scores ranged from 0 to 1; 0 indicated that the variable was almost absent, and 1 indicated that the variable was very strong. A teacher who has been teaching in a class will grade the samples differently in the second round. The mean of the six variables' coefficient of consistency is 0.82, and the consistency rate of the boundary category review is 86.5 per cent. To verify the stability of the model, the samples were divided into groups based on the type of activity, and the training, validation and test sets were 70%, 15% and 15% respectively. Table 2 shows the composition of the sample and reliability indicators. Table 2 shows that the coding consistency of task exercises and formative assessments is relatively high; that is, these two kinds of activities have clear data and rule foundations. The consistency among collaborative learning, social-emotional support, and ethical creation activities is slightly lower, and it is known that these activities have a broad context of explanation.

Table 2: Sample Composition and Coding Reliability

Activity Type	Sample Size	κ Consistency	ICC	Main Boundary Result
Content presentation	56	0.86	0.84	96.4% fully computable
Task practice	64	0.88	0.86	98.4% fully computable
Formative assessment	58	0.84	0.82	65.5% auxiliary computable
Collaborative learning	46	0.78	0.76	95.7% conditionally computable
Social-emotional support	38	0.74	0.72	100.0% human-led
Ethical-creative activities	50	0.76	0.74	98.0% human-led
Total / Mean	312	0.82	0.79	All four boundary categories covered

2.2 Computability Index and Hybrid Human-AI Decision Framework

After modularisation of the activity, an index is constructed in this paper. Let the vector of the i-th teaching activity unit be denoted by:

$$a_i = (d_i, r_i, t_i, k_i, s_i, e_i) \tag{1}$$

In the formula, d_i represents data observability, r_i represents rule dominance, t_i represents feedback immediacy, k_i represents situational complexity, s_i represents relationship sensitivity, and e_i represents ethical creation burden. All variables are standardized to the 0-1 interval. This representation preserves the multidimensional characteristics of teaching activities and provides input for the model to consider both technically tractable conditions and human judgment load simultaneously. Research on automatic evaluation has shown that text responses, objective questions, and regularized feedback can be well incorporated into model processing, but more detailed input-output frameworks are still needed for open-ended expressions and complex feedback [12]. This article uses six dimensional variables to jointly describe the activity boundary.

The computable index B_i is defined as:

$$B_i = 0.35d_i + 0.25r_i + 0.15t_i + 0.10(1 - k_i) + 0.10(1 - s_i) + 0.05(1 - e_i) \tag{2}$$

In the formula, the closer the value of B_i is to 1, the more suitable the activity is for independent processing by the computing system; The closer the value of B_i is to 0, the more the activity relies on the teacher's on-site judgment. The weight setting follows two principles: firstly, data observability and rule dominance are prerequisites for the stable operation of the computing system, with higher weights; Secondly, although the ethical creation load only accounts for 0.05 in numerical value, it still enters the risk gating stage and cannot be offset by weighted average. This setting reduces the risk of excessive automation of log rich or rule-based activities. For example, student psychological crisis identification may have high data observability, but its sensitivity to relationships and ethical risks determine that the system can only provide warnings and cannot replace teachers and professionals in making processing decisions.

Based on B_i and risk variables, this article divides teaching activities into four types of boundary intervals:

$$y_i = \begin{cases} C_1, B_i \geq 0.72, \max(s_i, e_i) < 0.45 \\ C_2, 0.60 \leq B_i < 0.72, \max(s_i, e_i) < 0.62 \\ C_3, 0.45 \leq B_i < 0.60, \max(s_i, e_i) < 0.82 \\ C_4, otherwise \end{cases} \quad (3)$$

In the formula, C_1 is fully computable, C_2 is auxiliary computable, C_3 is conditionally computable, and C_4 is human dominated. The four types of intervals correspond to different control rights configurations. C_1 allows the system to directly complete content push, objective practice feedback, and regular reminders; C_2 requires teachers to sample and review or set rule boundaries; C_3 can only use system suggestions after the teacher confirms the objectives and context; C_4 retains teacher led approach and limits AI output to information organization, material prompts, or risk clues. This classification incorporates the control concept from the complementary design research of teachers and intelligent systems [13], while also referencing the emphasis on teacher intervention timing in classroom scheduling tools [14].

Before inputting the model, this article performs two treatments on the variables. Firstly, merge the repeated behavioral evidence in the same activity description to prevent log intensive activities from gaining excessive weight. Secondly, risk markers should be set for activities that clearly rely on on-site relationships. As long as s_i or e_i exceeds 0.82, even if the candidate category reaches the C_2 threshold, it will be adjusted to C_4 or submitted for teacher review. This rule originates from the responsibility logic of educational scenarios: the system can help teachers see more evidence, but cannot directly provide final disposal in high-risk activities. After this processing, the model output not only includes the category, but also the main variables that trigger the category, making it easier for teachers to understand the basis for boundary judgment.

Index classification is used in the deployment phase, so it needs to work with the risk-gating and teacher-review mechanisms. Figure 3 is the hybrid human-machine boundary determination framework shown in Figure 3.

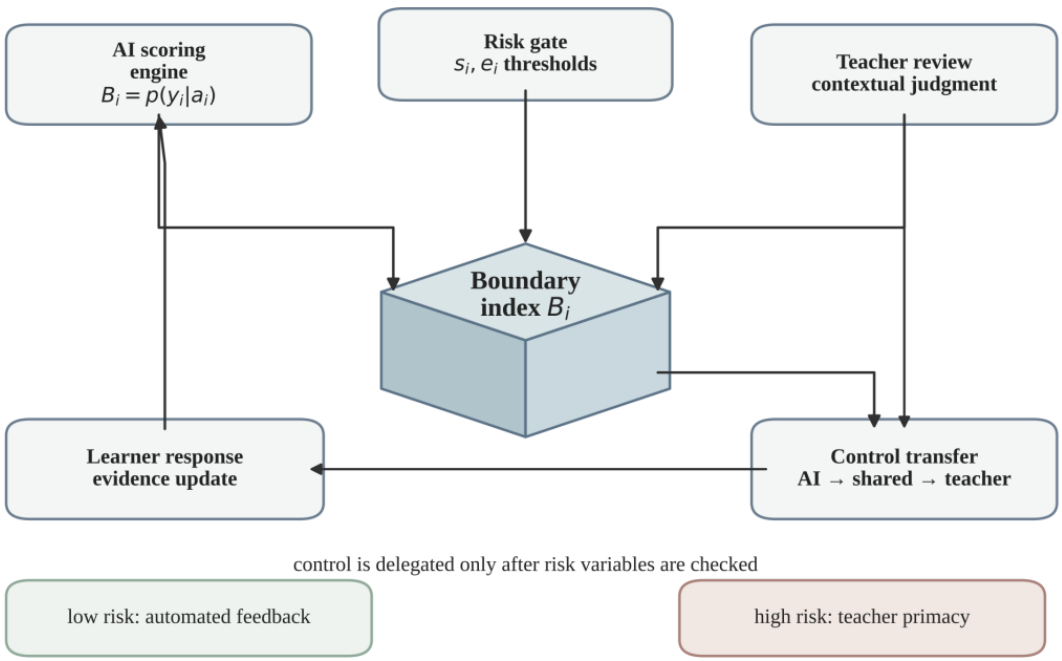

Figure 3: Hybrid Human-Machine Framework and Control Transfer Mechanism

As shown in Figure 3, the AI scoring engine outputs B1 and boundary candidate categories, and the risk gating reads S1 and E2 to compare them against the threshold. Based on the classroom goals, students' conditions, etc., adjust the control measures in the teacher's review module. The system will only provide understandable evidence, and the ultimate decision on disposal is still made by the teacher. The system will automatically carry out and save the logs for the C1 activity; for activity C2, a random check will be performed after execution; for the C3 activity, the system will provide suggestions first, and only then will the teacher confirm before entering the classroom; for the C4 activity, only material retrieval, risk alerts or record organisation will be provided by the system. Research on hybrid human-machine regulation has pointed out that the learning regulation power should be gradually transferred between AI and learners [15]; this paper embeds teacher review in boundary determination to prevent high-risk activities from crossing the teaching responsibility boundary due to high algorithm confidence.

To evaluate the total costs of the various control methods, a combined-decision objective function has been established in this paper.

$$\mathcal{L} = \lambda_1(1 - \mathrm{F1}) + \lambda_2 R + \lambda_3 H \tag{4}$$

In the formula, $\mathcal{L}$ represents the comprehensive cost, F1 represents the boundary recognition performance, R represents the residual risk, H represents the teacher's manual load, and λ_1 λ_2, and λ_3 are weight coefficients. This article sets the three weights as 0.50, 0.30, and 0.20 in the validation set, reflecting the principle of "prioritizing accurate identification while controlling risks and loads". Ethical research emphasizes that educational AI cannot be deployed solely based on efficiency, and responsibility, transparency, and fairness need to be incorporated into design goals; K-12 ethical research also reminds that high-risk student groups and child scenarios must be subject to stricter human intervention [16]. Therefore, the objective function of this article prioritizes an acceptable balance between accuracy, risk, and teacher burden.

2.3 Experimental Protocol and Evaluation Metrics

Set up four sets of comparison methods in the experiment. The first group is a rule baseline, classified only by the Bi threshold, and does not trigger risk gating. The second group is a semantic baseline; it classifies based on the activity description text and serves as a model for classification methods that use only semantics. The third group is the boundary index model, and its six dimensions and threshold values have been jointly judged. The fourth group is a hybrid human-machine framework in this paper that adds risk gating and teacher review rules based on the boundary index. The four ways use the same training and testing sets to avoid sample differences in the comparison results. Research has shown that AI feedback can be used to speed up the process of opening writing, feedback generation and discussion support; however, students still value individual understanding and emotional support from teachers' feedback [17]. Therefore, the aim of the comparative experiment is not only to evaluate the quality of automatically generated text but also to determine whether the model can distinguish between the categories of "auxiliary" and "teacher-led" guidance.

Limit the Scope of Parameter Settings. Threshold did not pass the test set callback; therefore, a threshold was set based on the quantile distribution of the four types of activities in the training set, and then residual risk was evaluated on the validation set. The semantic baseline does not participate in threshold learning; it only reads the activity descriptions and outputs the four types of labels to test whether the hypothesis that "sufficient text understanding can determine boundaries" is correct. The boundary index model has six dimensions, and the mixed framework adds risk gating and teacher review rules to the model. This paper will set out the

computable range and examine the problems of variability and incorrect application in the experiment. All the test results have a dimension for activity type to prevent the overall indicators from being masked by a low recall rate in social emotional support and ethical education activities.

Boundary Recognition Performance Metrics: Accuracy, Macro Average F1, and Weighted F1. Macro-average F1 is employed to determine if small-sample-category data are overlooked, and weighted F1 is used to show the general classification accuracy. It is as follows:

$$mathrmMacro-F1 = \frac{1}{4}\sum_{c=1}^{4} \frac{2P_c R_c}{P_c + R_c} \tag{5}$$

In the formula, P_c represents the precision of the c-class boundary, and R_c represents the recall of the c-class boundary. In addition to classification indicators, this article also records three deployment indicators: automation coverage, teacher review load, and residual risk. The automation coverage rate represents the proportion of activities that can be executed without the need for teacher confirmation item by item; Teacher review load refers to the number of activity units that require teacher confirmation for every 100 units; Residual risk refers to the proportion of activities with high ethical load or high relationship sensitivity that are mistakenly assigned to the automatic processing interval. Comparative studies on feedback sources indicate that AI feedback and peer feedback have different learning effects in writing tasks, and feedback quality, learner acceptance, and task nature need to be considered simultaneously [18]. Therefore, this article incorporates residual risks into the evaluation protocol to prevent the model from achieving surface accuracy on high-risk boundaries.

The four types of evidence in the test set output are: activity original summary, six dimensional variable scores, model prediction categories, and review categories. If the model prediction does not match the review category, record the reasons for the error, such as insufficient data, unclear rules, underestimation of relationship risk, underestimation of ethical creation load, and semantic misreading. Recordings can be made of the reasons for errors in judgment to avoid simply presenting a general accuracy rate. This framework is suitable for assessing the feasibility of teaching activities, does not directly measure students' learning outcomes, and is not an intervention-effect experiment in the classroom.

All the results are presented as the mean of three stratified samples, and the standard deviation is not retained. In addition, all rules that require teacher review are added to the framework in the form of "submitting judgments", and teachers can select the system classification and leave reasons. The above configuration will ensure that the model's results have an audit trail in production and also allocate space for threshold calibration for different schools, stages and subjects.

Threshold analysis scans with a step size of 0.05 in the range of 0.35-0.85 to observe the relationship among teacher load, risk and automation coverage. Error analysis will be carried out by case review, and for the activity units that were incorrectly classified in the test set, their original activity descriptions, six-dimensional variable values, model outputs, teacher's evaluation, and reasons for the error will be collected. This protocol combines activity distribution, model performance and threshold cost in the same validation chain, and corresponds to the boundary positions of teaching activities, recognition ability of the hybrid human-machine framework, as well as efficiency benefits and teaching risks under different risk constraints of intelligent system deployment.

3 Results and Discussion

3.1 Distribution of Computable Boundaries across Instructional Activity Categories

After finishing sample coding and boundary index calculation, the six types of teaching activities had different structures for the variables. The mean values of the six kinds of activities are shown in Figure 4.

Figure 4: Heat map of boundary variable distribution in teaching activities

As shown in Figure 4, the content presentation and task practice scores relatively high in the dimensions of data observability, rule dominance and real-time feedback, with ranges of 0.82-0.84, 0.78-0.82 and 0.71-0.76 respectively; at the same time, both relationship sensitivity and ethical creation load are below 0.20. The input and output of such activities are more stable, and the platform logs can cover most of the learning evidence. Content pushing and practice feedback from the algorithm are more convenient for teachers to review. Observability of formative assessment is 0.76; the proportion of rules is 0.69; and timeliness of feedback is 0.72. However, the situation complexity has risen to 0.46, and it is now considered to be in an intermediate state of "computable but interpretable". Collaborative learning, social-emotional support and ethical creation activities are not the same as the first three types of activities. The situation complexity of cooperative learning is 0.68 and the relationship sensitivity is 0.61. Although it is still at a data observability of 0.58, there is often a division of labour, unfairness, peer conflict and distribution of discourse power. Logs are not detailed enough to describe the quality of the interaction. The sensitivity to change of social-emotional support is 0.86, and the ethical construction load of ethical construction activities is 0.84; both are in the high human value load area. The digital education platform can observe changes in students' behaviour, and for emotions, understanding, trust-building and value judgment, teachers still need to deal with these based on on-site relationships. Although students have a high acceptance of generative artificial intelligence, privacy, accuracy and the impact on personal development are still the primary concerns [19]; teachers generally view generative tools as auxiliary resources and maintain a cautious attitude towards high-risk evaluation and relational teaching [20]. The distribution of the boundary in this paper is in line with the direction of these survey results.

There are continuous distributions of the boundary changes. Figure 5 is a computable exponential three-dimensional surface, as shown in Figure 5.

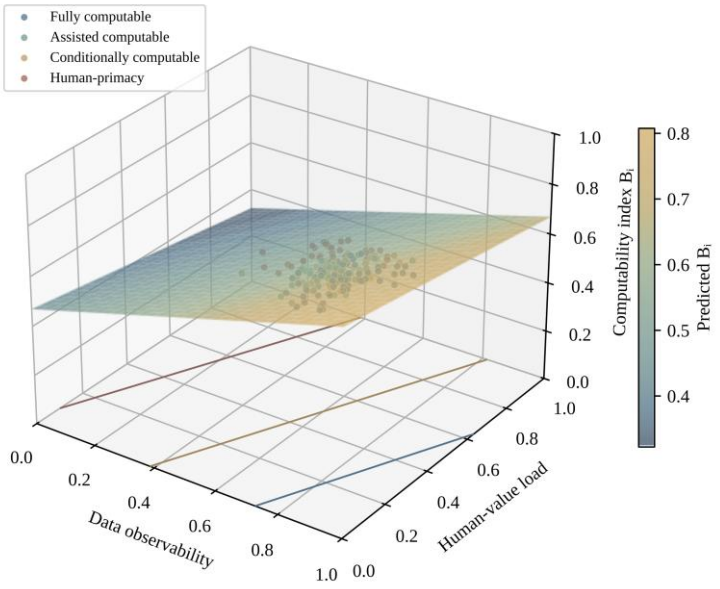

Figure 5: Computational Boundary 3D Surface and Activity Clustering

In Figure 5, an increase in data observability will push up B_i, while an increase in human value load will decrease B_i; There is a non-linear constraint relationship between the two. When the observability of data increases from 0.40 to 0.80 and the human value load remains below 0.30, B_i usually rises from around 0.55 to above 0.70, and the activity will transition from auxiliary computability to fully computable range. On the contrary, when the human value load exceeds 0.75, even if the data observability reaches 0.70, it is difficult for B1 to stably exceed 0.60, and activities will still be pushed back into the computable or human dominated range. This surface indicates that data can reduce recognition uncertainty, but cannot eliminate value and relational responsibility.

Based on the results of the category analysis, 96.4% of the activities in the content presentation were rated as fully computable, and only 3.6% were classified as auxiliary computable; 98.4% of the task exercises are fully computable, and only 1.6% are auxiliary computable. The range of formative assessment is relatively wide, and 34.5% are entirely measurable while 65.5% are indirectly observable. Therefore, although objective problem diagnosis, error induction and immediate feedback can be performed by the system, open interpretation, creative expression and individualised encouragement still need teachers to sample and review. 95.7 per cent of collaborative learning falls under conditional computability, and only 4.3 per cent falls under human dominance; all social and emotional support is human-dominated, and 98.0 per cent of ethical creation activities are human-dominated; only 2.0 per cent meet the computable conditions because of clear task rules. This distribution offers concrete examples of the judgments on AI's applications in content dissemination, assessment and feedback provision, and educational assistance, as well as indicating the limits of cooperation, emotions and ethics.

3.2 Boundary Identification Performance, Module Contributions, and Threshold Trade-Offs

Determine the distribution of different activities at the boundary, and then the model will test

how stable the boundary recognition is. Table 3 shows the general results of the four methods on the test set, as follows: The accuracy of the rule baseline is 0.771 and the weighted F1 is 0.745. Most of the errors occur in the formative evaluation and collaborative learning phases. The accuracy of the semantic baseline has reached 0.807 and the weighted F1 is 0.780. The activity description text may offer some reference points, but the risk relationships and ethical construction responsibilities are generally not mentioned. The boundary index model increases the weighted F1 to 0.812 and lowers the residual risk from 9.8% for the rule baseline to 5.2%. The hybrid human-machine framework performed best and had the following results: accuracy of 0.873, macro-average F1 of 0.821, weighted F1 of 0.845 and reduced residual risk of 2.9 per cent. Weighted F1 has increased by 10.0 percentage points over the baseline rule, and by 6.5 percentage points over the semantic baseline.

Table 3: General performance of the above boundary recognition methods

Method	Accuracy	Macro-F1	Weighted F1	Residual Risk	Teacher Review Load
Rule baseline	0.771	0.713	0.745	9.8%	21.4/100
Semantic baseline	0.807	0.748	0.780	7.1%	28.7/100
Boundary index model	0.836	0.787	0.812	5.2%	35.5/100
Hybrid human-AI framework	0.873	0.821	0.845	2.9%	42.6/100

In terms of the contribution of modules, adding the six-dimensional variables can better differentiate between C2 and C3, and the risk gating primarily reduces the probability of classifying C4 as C1 or C2. The rule baseline has high accuracy for content display and task practice, but there is a lack of subdivision in formative assessment; Semantic baselines can identify textual cues such as "discussion", "appeasement" and "value judgment", but are unstable in describing shorter activities. The two strengths of the hybrid model are: first, using variable indices to select the conditions for calculation; and second, employing si and ei to evaluate the risks of teaching responsibility. Combining the two will not only provide coverage but also shift the high-risk misjudgments to the teacher's review list.

Figure 6 shows the difference of the model from the perspective of activity types, as shown in Figure 6.

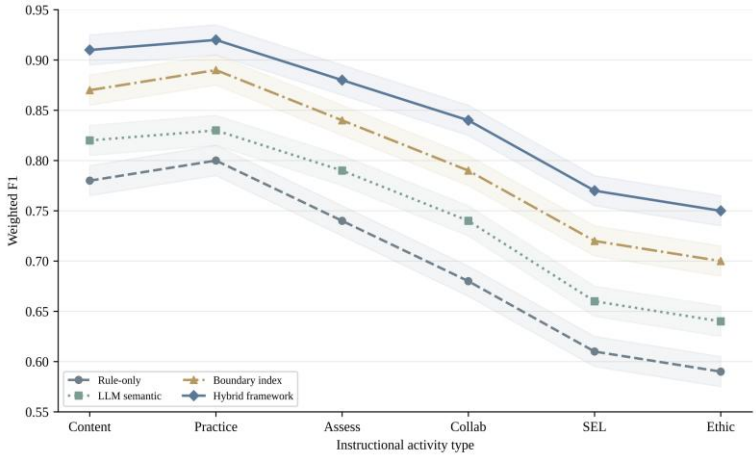

Figure 6: Boundary Recognition Performance of Different Model Configurations

As shown in Figure 6, the hybrid human-machine framework has superior performance among the other methods for all six types of activities, with weighted F1 scores of 0.91 for content presentation and 0.92 for task practice, 0.88 for formative evaluation, and 0.84 for collaborative learning. The F1 scores of the social-emotional support and ethical creation activities are relatively low at 0.77 and 0.75, respectively, but still higher than the baseline values of 0.61 and 0.59. Risk gating and teacher review rules are more stringent for high human value load activities. The base rules have a high accuracy rate for low-risk behaviour; however, when dealing with activities that seem computable but are in fact subject to teacher discretion, there is an increase in misjudgments. Semantic baselines can use activity descriptions but lack interpretable variable constraints; when words such as "feedback", "suggestion" and "evaluation" appear in the description, it is easy to mistake the task for auxiliary computability. The six reduced dimensions of the boundary index model reduce misjudgments, and a hybrid framework sets thresholds for s_i and e_i in high-risk activities.

Threshold Selection Will Affect the Deployment Results of the System. Figure 7 shows the impact of the control transfer threshold τ on accuracy, teacher workload, residual risk and automation coverage, as shown in Figure 7.

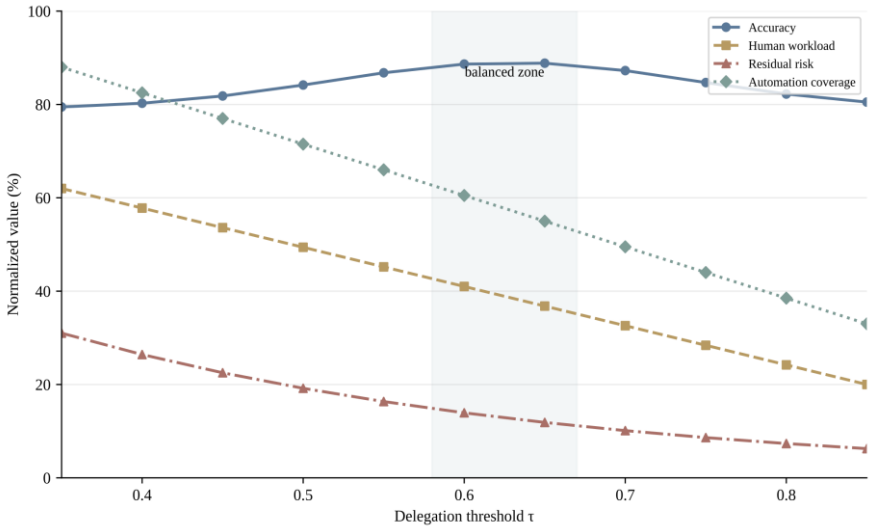

Figure 7: Impact of Human-Machine Collaboration Threshold on Accuracy, Manual Load and Risk

In Figure 7, when τ increases from 0.35 to 0.55, the residual risk decreases from 31.0% to approximately 12.2%, but the automation coverage also decreases from 88.0% to around 66.0%. When τ continues to rise to 0.65, the accuracy reaches the highest range, about 88.5%, the residual risk drops to 6.5%, and the teacher review load is about 37.0 activity units/100. After τ exceeds 0.70, the residual risk continues to decrease, but the automation coverage rate drops below 50%, and the teacher workload increases significantly. Overall, the range of 0.58-0.67 is relatively balanced, which can control residual risks within an acceptable range while retaining about half of the automation benefits.

This change suggests that the performance of models in education should be explained in conjunction with the direction of the misjudgment. In the study of computable boundaries, assuming that human-dominated activities are fully computable will incur a high cost for teacher sampling compared with submitting only fully computable activities.

The indices of convenience are the same. The mixed framework requires teachers to review 42.6 per hundred activity units, which is higher than the 35.5 in the boundary index model, but

the residual risk is reduced from 5.2% to 2.9%. To reduce the occurrence of a single high-risk misjudgment, the added review work is about 3.1 activity units and still within the reasonable limit for teachers. Teacher review limits the advantages of automation to a certain extent within the scope of teaching duties.

The Threshold result is related to the deployment method of the school. If schools aim for the highest level of automation coverage, the system will push some conditional computable activities towards automatic processing, and the main risks will be in student emotion recognition, peer conflict prompts, and value writing feedback. If the threshold set by the school is too high, the amount of work for teachers will be close to that of manual review, and the efficiency gain of intelligent systems will be reduced. A reasonable way is to set different thresholds according to the type of activity: a low threshold can be used for content presentation and task practice, a medium threshold can be employed for formative assessment, and high thresholds or direct instruction by teachers may be needed for collaborative learning, social-emotional support, and ethical creation activities. Research on generative AI education indicates that technology empowerment should be carried out in tandem with adjustments to teaching methods and theoretical foundations, and a systematic review of technology availability shows that the advantages of tools can only be converted into stable educational values through the coordination of tasks, learners and teachers. The threshold will be used in this paper to make a quantitative match.

3.3 Case Trajectories, Error Sources, and Deployment Implications

The case trajectory can show the particular way the model output enters the teaching explanation. Figure 8 selects six representative activities to draw the case trajectory, as shown in Figure 8.

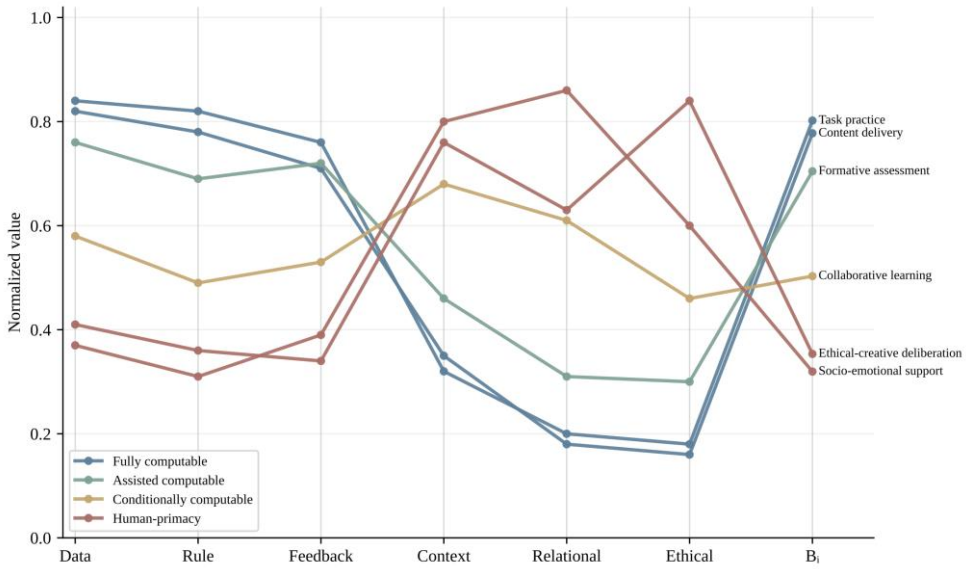

Figure 8: Case Trajectories of Typical Teaching Activities in Four Boundary Intervals

In Figure 8, the curves of content presentation and task practice remain high in the dimensions of data, rules, and feedback, and low in the dimensions of relationships and ethics, with final values of 0.78 and 0.81, respectively. The B_i of formative assessment is about 0.71, which is close to the fully computable threshold, but due to the uncertainty of open interpretation and individual feedback, it is classified as auxiliary computable. The B_i of collaborative learning is about 0.53, and both situational complexity and relationship sensitivity

increase. The model classifies it as conditionally computable. The B_i values of social emotional support and ethical creation activities are both below 0.35, and the relationship or ethical variables exceed 0.80, thus entering the human dominated range. Figure 8 decomposes the boundary judgments of each case into variable levels, which teachers can use to verify system control suggestions.

According to the error analysis, a total of 14 activity units in the test set were incorrectly recognised by the hybrid framework. The three origins of the errors are as follows. The first kind is variable underestimation due to semantic compression, at a rate of 42.9%. For example, 'peer evaluation after group presentation' is close to traditional evaluation in text, but in real classrooms, it may involve peer relationships, fear of the teacher, and issues of fair participation; thus, the model is likely to have a low rating for relationship sensitivity. Secondly, too much emphasis is placed on the rules of behaviour (35.7%). For example, although the criteria for evaluating project-based learning are clear, the original content and ethical implications, as well as practical applicability, of the student's plans still need to be judged by teachers. The third is a lack of student status information in the activity description, accounting for 21.4%. For example, 'sending reminders to students who have missed consecutive assignments' may appear to be a regular management task, but if there are psychological pressures or family problems for the students, the system's automatic reminders may add more stress. The SWOT analysis of ChatGPT for education has also identified that, at the same time, efficiency, accessibility and personalised support need to be achieved; otherwise, risks such as bias, over-reliance, and weakened teaching relationships may arise [23]. The error distribution requires the boundary model to keep the mechanism of "submitting to the teacher when information is insufficient".

The three directions for the deployment of the above research results are as follows: First of all, low-risk and high-rule activities can be prioritized for deployment in intelligent teaching systems, such as resource push, objective practice feedback, routine error statistics and learning reminders. These activities have a high level of B1, and teachers generally only carry out goal-setting and sample-inspection. Second, the activities in the middle stage should adopt a "system-generated suggestion, teacher-selected execution" model for formative assessment, open-ended homework feedback and collaborative task arrangement. Thirdly, the high sensitivity of the relationship and the need for ethical creation activities require teacher leadership; therefore, the system can only provide material retrieval, process recording and risk warning functions. Based on a systematic review of teacher professional development, most of the AI research has focused on teaching applications rather than how to help teachers learn to use, assess and modify AI recommendations [24]. A computable boundary has been established in the teacher training programme to help teachers decide when to employ artificial intelligence (AI), how to assess the results of using AI, and at what time to disregard AI recommendations.

The framework of this paper elaborates on the concept of "AI as a teaching assistant" and proposes deployment rules. Complete Computable Activities can enhance efficiency, and auxiliary computable activities can reduce the load of repeated feedback. Teacher supervision is required for conditional computable activities; otherwise, human-led activities promote community education and cultivate creativity. Research on the interaction of learning analytics and artificial intelligence suggests that future educational technology will be a type of hybrid intelligence, and algorithms will process stable-to-be-expressed evidence, while humans will take charge of setting goals, interpreting meaning, and bearing responsibility. The results of this paper are in line with the above and apply hybrid intelligence to teaching activity units and boundary thresholds. This study has only been using sample-encoded data so far and has not been observing the learning effect in a real classroom. In the future, it can be added to the school platform and other improvements, and continuously adjust the thresholds at different levels, courses and classes, and observe how changes in teachers' review behaviour affect students'

learning experiences.

4 Conclusion

This paper puts forward a hybrid human-machine framework to focus on the computable scope of teaching activities in the context of digital education, uses activity units as the object, six-dimensional variables as the basis, and risk gating and teacher review as constraints. Research has shown that the ability of teaching activities to be scored depends on whether the platform can save the data; also, there are issues such as unclear rules and inconsistent feedback, as well as a lack of relationships, emotions, ethics and creative judgment in the activities.

(1) Firstly, this article organizes 312 teaching activity units into programmable objects and establishes six indicators: data observability, rule dominance, feedback immediacy, situational complexity, relationship sensitivity, and ethical creation load, enabling the "computable boundary" to have a verifiable data path.

(2) Secondly, the weighted F1 of the hybrid human-machine framework on the test set reached 0.845, which is 10.0 percentage points higher than the rule baseline and reduces residual risk to 2.9%. The results indicate that content presentation, task practice, and partial formative evaluation are suitable for limited automation, collaborative learning requires the conditional use of AI, and social emotional support and ethical creation activities should maintain teacher leadership.

(3) Thirdly, this article still belongs to the research of sample coding and has not directly tested the long-term learning effectiveness in real classrooms. Future research can integrate real teaching data into different stages and subject platforms, continuously calibrate thresholds, and examine the relationship between teacher review behavior, student acceptance, and learning outcomes.

About the Author

Jiayi Li was born in Heze, Shandong Province, China in 2003. She graduated from Shandong University of Political Science and Law in China and obtained a bachelor's degree. She is now a student at the Academy for Applied Policy Studies and Education Futures and the Education University of Hong Kong. Her primary direction of study is Emerging Technology for Future Workforces.

References

[1] Zawacki-Richter, O., Marín, V. I., Bond, M., et al. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? International Journal of Educational Technology in Higher Education, 16, 39.

[2] Ouyang, F., Jiao, P. (2021). Artificial intelligence in education: The three paradigms. Computers and Education: Artificial Intelligence, 2, 100020.

[3] Kasneci, E., Sessler, K., Küchemann, S., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, 102274.

[4] Wang, S., Wang, F., Zhu, Z., et al. (2024). Artificial intelligence in education: A

systematic literature review. Expert Systems with Applications, 252, 124167.

[5] Molenaar, I. (2022). Towards hybrid human-AI learning technologies. European Journal of Education, 57(4), 632–645.

[6] Alfredo, R., Echeverria, V., Jin, Y., et al. (2024). Human-centred learning analytics and AI in education: A systematic literature review. Computers and Education: Artificial Intelligence, 6, 100215.

[7] Crompton, H., Burke, D. (2023). Artificial intelligence in higher education: The state of the field. International Journal of Educational Technology in Higher Education, 20, 22.

[8] Martin, F., Zhuang, M., Schaefer, D. (2024). Systematic review of research on artificial intelligence in K-12 education (2017–2022). Computers and Education: Artificial Intelligence, 6, 100195.

[9] Holmes, W., Porayska-Pomsta, K., Holstein, K., et al. (2022). Ethics of AI in education: Towards a community-wide framework. International Journal of Artificial Intelligence in Education, 32, 504–526.

[10] Miao, F., Holmes, W. (2023). Guidance for Generative AI in Education and Research. Paris: UNESCO.

[11] Maghsudi, S., Lan, A., Xu, J., et al. (2021). Personalized education in the artificial intelligence era: What to expect next. IEEE Signal Processing Magazine, 38(3), 37–50.

[12] Gao, R., Merzdorf, H. E., Anwar, S., et al. (2024). Automatic assessment of text-based responses in post-secondary education: A systematic review. Computers and Education: Artificial Intelligence, 6, 100206.

[13] Holstein, K., McLaren, B. M., Aleven, V. (2019). Designing for complementarity: Teacher and student needs for orchestration support in AI-enhanced classrooms. In Artificial Intelligence in Education. Cham: Springer, 157–171.

[14] Holstein, K., McLaren, B. M., Aleven, V. (2019). Co-designing a real-time classroom orchestration tool to support teacher–AI complementarity. Journal of Learning Analytics, 6(2), 27–52.

[15] Molenaar, I. (2022). The concept of hybrid human-AI regulation: Exemplifying how to support young learners' self-regulated learning. Computers and Education: Artificial Intelligence, 3, 100070.

[16] Akgun, S., Greenhow, C. (2022). Artificial intelligence in education: Addressing ethical challenges in K-12 settings. AI and Ethics, 2(3), 431–440.

[17] Steiss, J., Tate, T., Graham, S., et al. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. Learning and Instruction, 91, 101894.

[18] Banihashem, S. K., Kerman, N. T., Noroozi, O., et al. (2024). Feedback sources in essay writing: Peer-generated or AI-generated feedback? International Journal of Educational Technology in Higher Education, 21, 23.

[19] Chan, C. K. Y., Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. International Journal of Educational Technology in Higher Education, 20, 43.

[20] Lee, D., Arnold, M., Srivastava, A., et al. (2024). The impact of generative AI on higher education learning and teaching: A study of educators' perspectives. Computers and Education: Artificial Intelligence, 6, 100221.

[21] Noroozi, O., Soleimani, S., Farrokhnia, M., et al. (2024). Generative AI in education: Pedagogical, theoretical, and methodological perspectives. International Journal of Technology in Education, 7(3), 373–385.

[22] Wang, N., Wang, X., Su, Y. S. (2024). Critical analysis of the technological affordances, challenges and future directions of generative AI in education: A systematic review. Asia Pacific Journal of Education, 44(1), 139–155.

[23] Farrokhnia, M., Banihashem, S. K., Noroozi, O., et al. (2024). A SWOT analysis of ChatGPT: Implications for educational practice and research. Innovations in Education and Teaching International, 61(3), 460–474.

[24] Tan, X., Cheng, G., Ling, M. H. (2025). Artificial intelligence in teaching and teacher professional development: A systematic review. Computers and Education: Artificial Intelligence, 8, 100355.

[25] Cukurova, M. (2025). The interplay of learning, analytics and artificial intelligence in education: A vision for hybrid intelligence. British Journal of Educational Technology, 56(2), 469–488.