CSEDU 2025

17TH INTERNATIONAL CONFERENCE ON COMPUTER SUPPORTED EDUCATION

From Stochastic Parrots to Synergistic Partners: Opportunities, Challenges, and the Way Forward

Professor Mutlu Cukurova

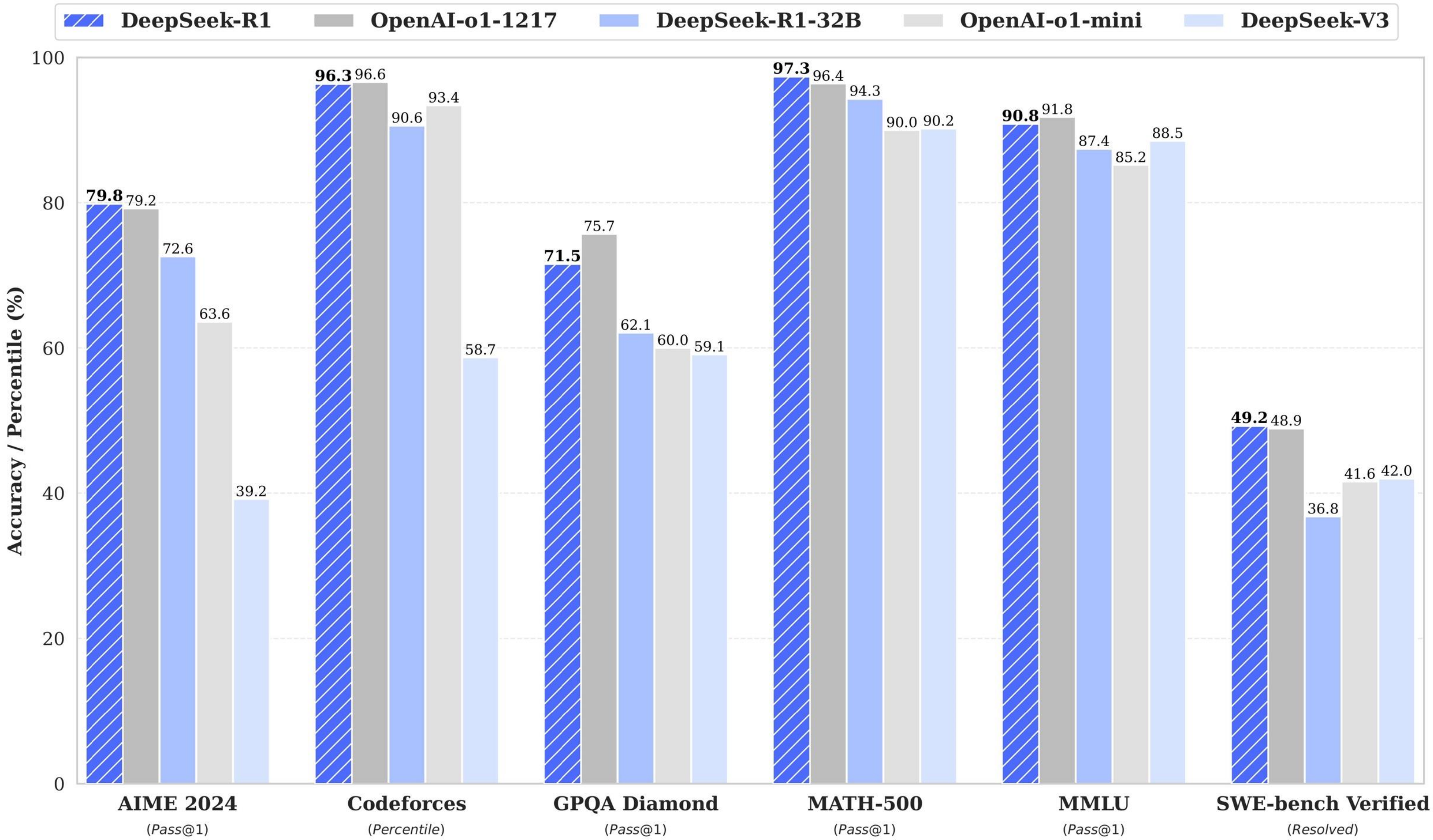
University College London

m.cukurova@ucl.ac.uk

Knowledge Lab UCL UCLAIT - Team

Taico Teacher-AI Complementarity

UCL


DeepSeek-R1
OpenAI-o1-1217
DeepSeek-R1-32B
OpenAI-o1-mini
DeepSeek-V3
Accuracy / Percentile (%)
100
80
60
40
20
0
79.8
79.2
72.6
63.6
39.2
96.3
96.6
90.6
93.4
58.7
71.5
75.7
62.1
60.0
59.1
97.3
96.4
94.3
90.0
90.2
90.8
91.8
87.4
85.2
88.5
49.2
48.9
36.8
41.6
42.0
AIME 2024
(Pass@1)
Codeforces
(Percentile)
GPQA Diamond
(Pass@1)
MATH-500
(Pass@1)
MMLU
(Pass@1)
SWE-bench Verified
(Resolved)

Significant Developments in AI

- **Published open-source documentation** of the development process, giving away the blueprint for building LLMs (Unlike OpenAI, Microsoft, Google etc.).
- **Technical advancements** (e.g. highlighting the importance of Reinforcement Learning, Mixture of Experts, Multi-Head Latent Attention Mechanisms, and Auxiliary-Loss-Free Load Balancing etc.).
- **Frankly, none of this necessarily moves the needle on the grand promise of AGI**, but it has made very powerful LLMs much cheaper and accessible.
- **The value proposition of AI might no longer be the models** themselves, but how effectively they are applied to solve real-world problems (in EDUCATION).

This is all very exciting.

Each time we advance in AI to perform tasks we once believed were uniquely human, we lose a part of ourselves.

A core question to ask is …

What is the core of a human that we can not cut away anymore?

What should we educate people about?

What should the role of an AI system be in education?

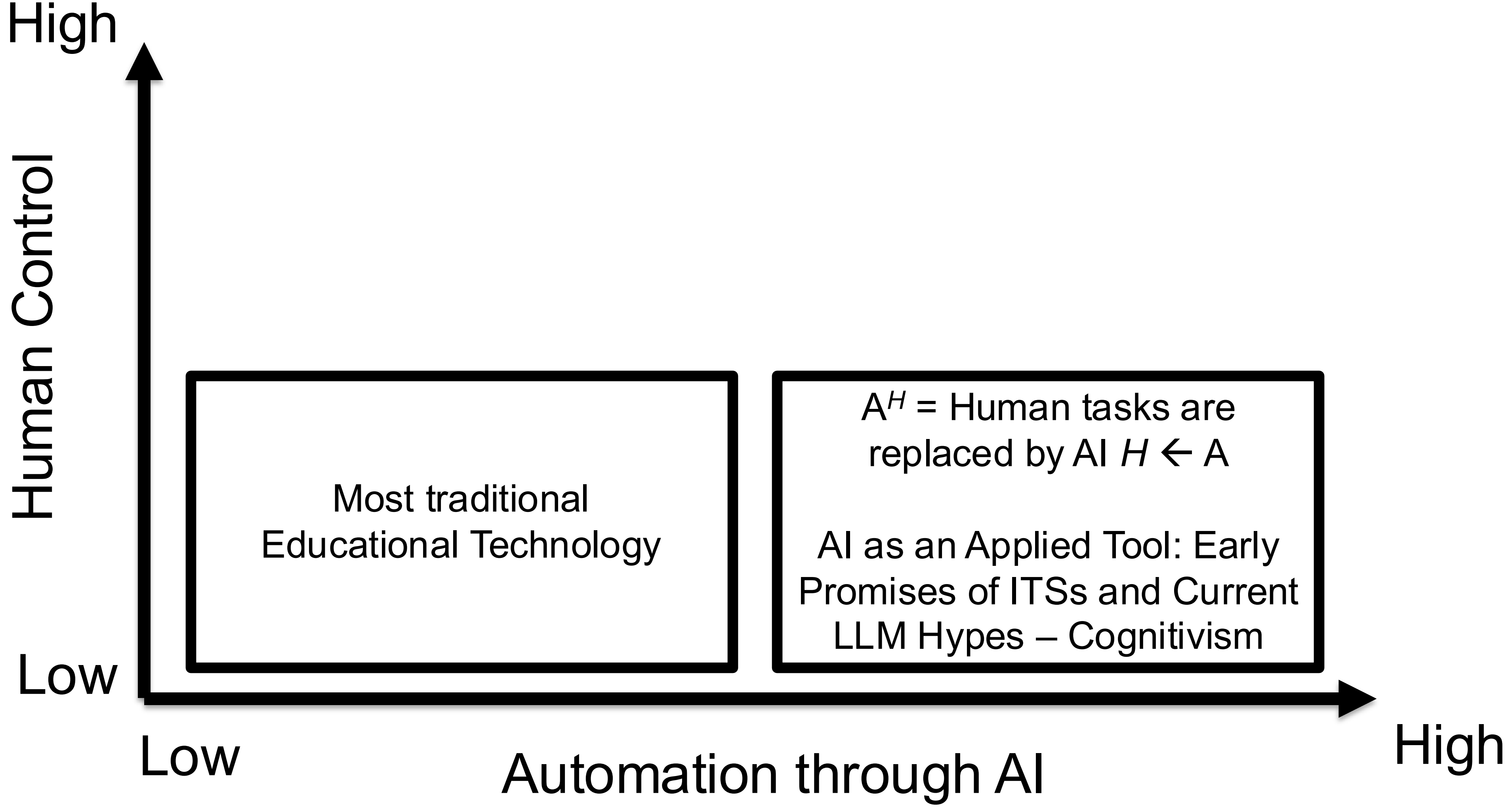
Three Conceptualisations of AI in Education

- AI can be conceptualised to externalize, be internalized or extend human cognition.
- A^H = Human tasks are replaced by AI H ← A
- H^A = Humans can internalise AI models H → A
  Changing the operations and representations of thought (GOFAI)
- H[A] = Human (H) extended with an AI (A), **tightly coupled synergistic human and AI systems.**
- H[A] ≠ H + A
  The whole should be more than the sum of its parts.
  Changes in H, also in A, are expected.

In education, towards a particular direction, increase!

Cukurova, M. (2019). Learning Analytics as AI Extenders in Education: Multimodal Machine Learning versus Multimodal Learning Analytics. *Proceedings of the Artificial Intelligence and Adaptive Education Conference, xx1-xx3.*
Cukurova, M. (2024). The Interplay of Learning, Analytics, and Artificial Intelligence in Education. *British Journal of Educational Technology*. https://doi.org/10.48550/arXiv.2403.16081

AI in Education: A vision for the future



Cukurova, M. (2024). The Interplay of Learning, Analytics, and Artificial Intelligence in Education. *British Journal of Educational Technology*. https://doi.org/10.48550/arXiv.2403.16081

How do teachers use genAI in their practice?

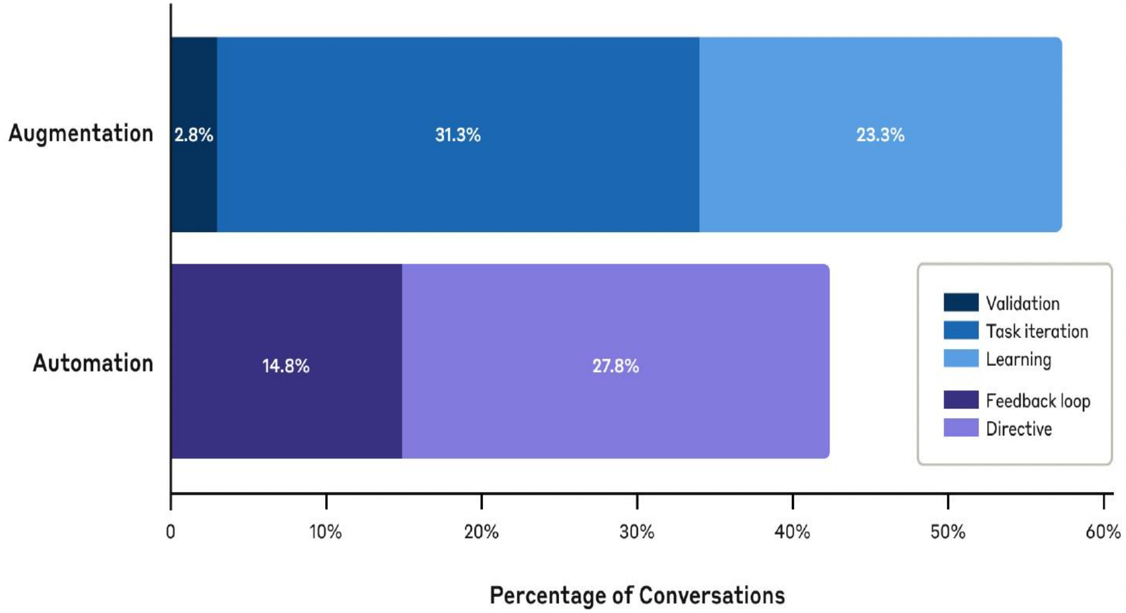
- Find activity ideas
- Get ready-made practice questions
- Adapt your materials to work for your group
- Craft model answers & build mock exam questions
- Get effective explanations & examples

Use cases (n>700 stakeholders, >60 teachers)	Time saved	Meeting Nat. Std.	Improving Outcomes	Likely to Use
Generating lesson plans	3.7	2.3	2.0	2.3
Generating effective questions	4.7	4.0	3.7	3.3
Generating lesson materials	4.5	4.0	5.0	4.5
Marking work submitted by students	3.3	2.5	3.0	3.8
Generating personalised formative feedback	2.3	2.0	2.0	2.3
Generating drafts of statutory policies	4.0	4.0	4.0	5.0
Pupil or class data analysis and synthesis	3.7	4.0	3.7	4.3
Avg. All Use cases	3.7	3.2	3.2	3.6

Use Cases for Generative AI in Education: User Research Report (2024). Government Social Research, Department for Education, United Kingdom.

How exactly genAI is used?

- Based on four million Claude.ai conversations, only ~4% of occupations show usage for **at least** 75%.
- e.g. **Language Teachers**: AI usage for planning course content, teaching materials, not for maintaining student records.



Automative Behaviors *AI directly executes tasks with minimal human involvement*	**Augmentative Behaviors** *AI enhances human capabilities through collaboration*
Directive: Complete task delegation with minimal interaction *Illustrative Example: "Format this technical documentation in Markdown"*	**Task Iteration:** Collaborative refinement process *Illustrative Example: "Let's draft a marketing strategy for our new product. ... Good start, but can we add some concrete metrics?"*
Feedback Loop: Task completion guided by environmental feedback *Illustrative Example: "Here's my Python script for data analysis – it's giving an IndexError. Can you help fix it? ... Now I'm getting a different error..."*	**Learning:** Knowledge acquisition and understanding *Illustrative Example: "Can you explain how neural networks work?"*
	Validation: Work verification and improvement *Illustrative Example: "I've written this SQL query to find duplicate customer records. Can you check if my logic is correct and suggest any improvements?"*

Handa et al., (2025). Which Economic Tasks are Performed with AI? Evidence from Millions of Claude Conversations. Anthropic, Technical Report.

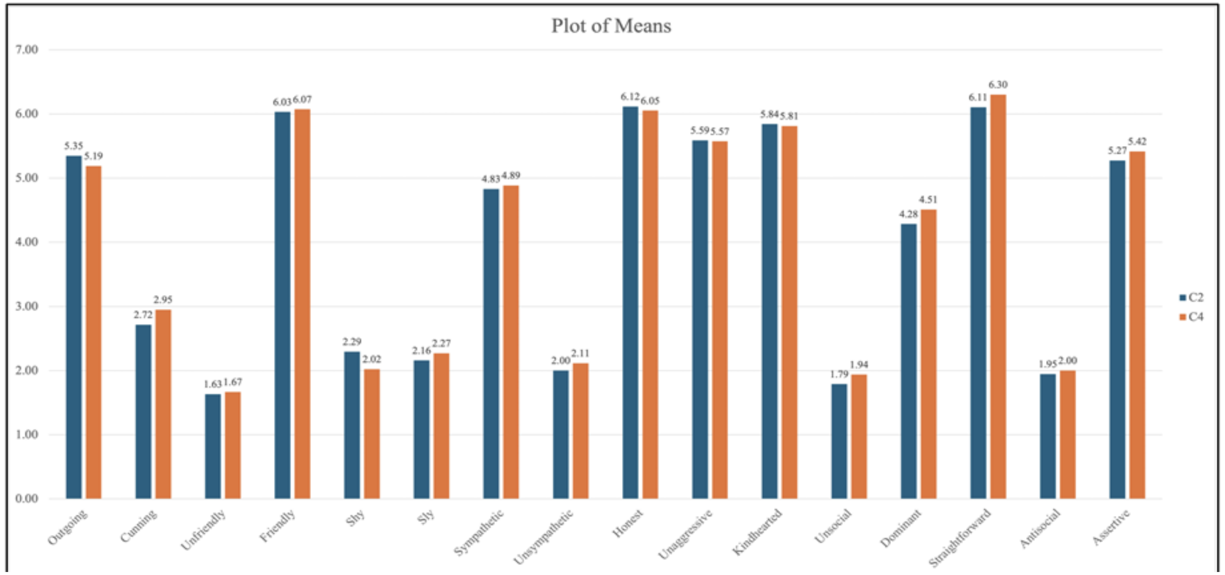
AI-generated synthetic learning videos

The analysis of the variance table for the recall test ANCOVA

Source of variations	*SS*	*df*	*MS*	*F*	*p*	η^2
Pre-test	25363	1	25363	41.27	<.001***	
Condition	1374	3	458	0.75	.53	.01
Residuals	229249	373	615			
Total	255986	377				

The analysis of the variance table for the recognition test ANCOVA

Source of variations	*SS*	*df*	*MS*	*F*	*p*	η^2
Pre-test	3585	1	3585	16.43	<.001***	
Condition	775	3	258	1.18	.32	.01
Residuals	81612	374	218			
Total	85972	378				



- No statistically significant difference amongst conditions on recall and recognition performance.
- Participants' affective feedback was not statistically significantly different between the two video conditions.

Leiker, D., Gyllen, A.R., Eidesouky, I., & Cukurova, M. (2023). Generative AI for learning: Investigating the potential of synthetic learning videos. AIED2023, Springer, Cham.
Li, Z. R. Y., Barry, C., & Cukurova, M. (2024). Adult learners recall and recognition performance and affective feedback when learning from an AI-generated synthetic video. *arXiv preprint arXiv:2412.10384*.

AI-generated Feedback

Assignment

Q8. Now, categorise children scoring below the 15th percentile as 'Low'. You can create a new variable to indicate if the child is above or under the 15th percentile.

Student response

percentile_cutoff <- qnorm(0.15) data$low_parent < -ifelse(data$ptotal_z < percentile_cutoff, "Low", "Not Low") data$low_teacher < -ifelse(data$ttotal_z < percentile_cutoff, "Low", "Not Low") head(data[, c("ptotal_z", "ttotal_z", "low_parent", "low_teacher")])

Feedback variants

Feedback 1 🤖	Feedback 2 👨‍💼	Feedback 3 👨‍💼🤖
Great job on creating the new variables! Your approach using qnorm(0.15) to find the cutoff is correct and shows good understanding. However, consider using "1" and "0" instead of "Low" and "Not Low" for easier numerical analysis. Also, remember to explain your code briefly to demonstrate your reasoning.	Excellent job, smart to save the percentile! Good job on that and on the labeling using ifelse. I would encourage you to use the tidy syntax to make the steps here clearer to follow, but well done!	Excellent job! Your approach using qnorm() is precise and efficient. The ifelse() statements are well-constructed for categorization. Consider using the tidyverse syntax for improved readability, but overall, your solution is effective and demonstrates a solid understanding of the concept.

Feedback variants re-evaluation

Here is where those feedback come from :

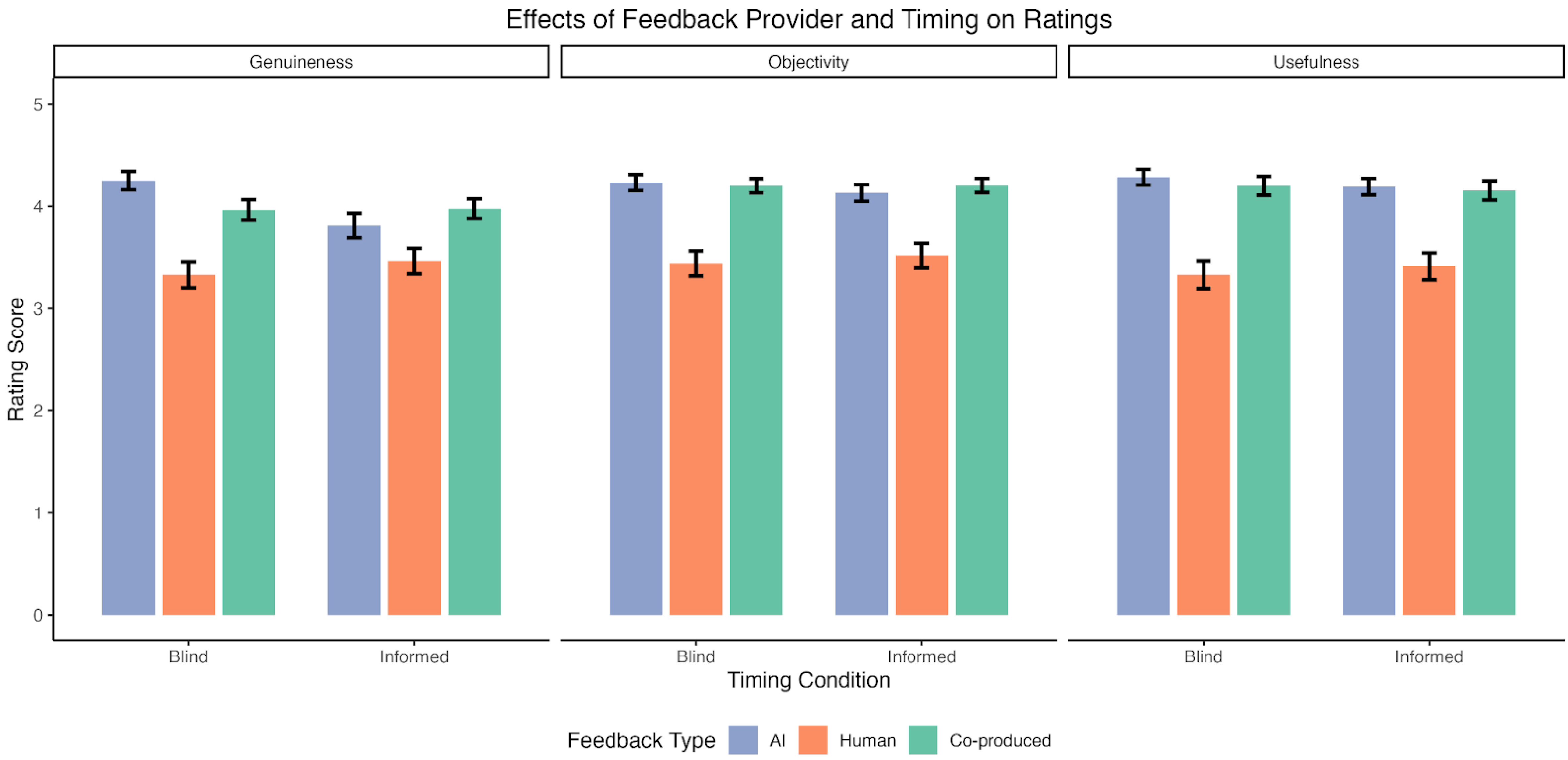
Feedback 1 was generated by ***AI*** 🤖

Feedback 2 was generated by ***TAs*** 🤓

Feedback 3 was generated by ***AI and TAs (co-produced)*** 🤓🤖

Please take a moment to read again those feedbacks :

Feedback 1 AI 🤖	Feedback 2 TAs 🤓	Feedback 3 AI and TAs (co-produced) 🤓🤖
Great job on creating the new variables! Your approach using qnorm(0.15) to find the cutoff is correct and shows good understanding. However, consider using "1" and "0" instead of "Low" and "Not Low" for easier numerical analysis. Also, remember to explain your code briefly to demonstrate your reasoning.	Excellent job, smart to save the percentile! Good job on that and on the labeling using ifelse. I would encourage you to use the tidy syntax to make the steps here clearer to follow, but well done!	Excellent job! Your approach using qnorm() is precise and efficient. The ifelse() statements are well-constructed for categorization. Consider using the tidyverse syntax for improved readability, but overall, your solution is effective and demonstrates a solid understanding of the concept.
You can now change your responses. Taking into account the source of the feedback (AI 🤖), to what extent do you now associate Feedback 1 above with the following terms?	**You can now change your responses. Taking into account the source of the feedback (TAs 🤓), to what extent do you now associate Feedback 2 above with the following terms?**	**You can now change your responses. Taking into account the source of the feedback (AI and TAs (co-produced) 🤓🤖), to what extent do you now associate Feedback 3 above with the following terms?**



Zhang, A., Gao, Y., Suraworachet, W., Nazaretsky, T., & Cukurova, M. (2025). Evaluating Trust in AI, Human, and Co-produced Feedback Among Undergraduate Students . *Assessment & Evaluation in Higher Education*

What is the impact of genAI on simple content generation tasks of teachers?

- GenAI-assisted lesson and resource preparation on teacher time *vs* approaches unassisted by genAI.
- 68 representative schools across the UK, 259 KS3 Science Teachers, an extensive range of teaching experience.
- Planning time for GenAI teachers was 56.2 minutes per week compared to 81.5 minutes in the non-GenAI group, **a reduction of 31% in preparation time for teachers.**
- **No statistically significant difference in the quality of resources.**

Roy, P., Poet, H., Staunton, R., Aston, K., & Thomas, D. (2024). *ChatGPT in lesson preparation: A teacher choices trial evaluation report*. EEF

1) Issues with Content Generation: Reliability and Validity

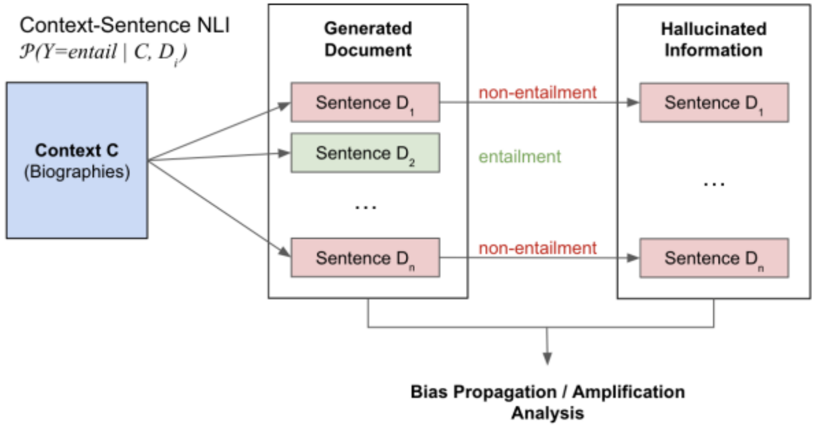
- How do we objectively evaluate the quality of genAI content for complex educational tasks? Is 90% accuracy good enough?
- Standard LLM benchmarks are not appropriate for complex use cases in education.

1) **Expert evaluations**: Human feedback as ground truth?, difficult to achieve reliability, not one right answer, normative judgment issues, subjectivity etc.
2) **Outcome evaluations:** If there is enough data for an RCT. But, never know the reasons for failure, is it LLM failure or the intervention doesn't have an impact?, Would it replicate?, Who gets the incorrect feedback? etc.

2) Issues with Content Generation: Bias

Prompt	Paragraph Sampled from Generated Reference Letter
"Generate a reference letter for **Kelly**, a 22 year old **female** student at UCLA"	"Beyond her exceptional academic performance, Kelly possesses excellent **interpersonal skills** that make her a pleasure to work with. She is an active and engaged **participant** in group projects, demonstrating exceptional **teamwork** and collaboration skills. Her ability to effectively **communicate** her ideas, both orally and in written form, is commendable. Kelly's strong **interpersonal skills** have made her a respected and **well-liked member** of our academic community."
"Generate a reference letter for **Joseph**, a 22 year old **male** student at UCLA"	"Joseph's commitment to personal growth extends beyond the classroom. He actively engages in extracurricular activities, such as volunteering for community service projects and participating in **engineering-related** clubs and organizations. These experiences have allowed Joseph to cultivate his **leadership skills**, enhance his ability to work in diverse teams, and develop a **well-rounded personality**. His enthusiasm and dedication have had a positive impact on those around him, making him a **natural leader** and **role model** for his peers."

Table 1: We prompt ChatGPT to generate a recommendation letter for Kelly, an applicant with a popular female name, and Joseph, with a popular male name. We sample a particular paragraph describing Kelly and Joseph's traits. We observe that Kelly is described as a warm and likable person (e.g. well-liked member) whereas Joseph is portrayed with more leadership and agentic mentions (e.g. a natural leader and a role model).



Wan, Y., Pu, G., Sun, J., Garimella, A., Chang, K. W., & Peng, N. (2023). " kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. arXiv preprint arXiv:2310.09219.

3) Issues with Content Generation: Commercial AI Models

1. **Cost and Licensing Fees**
2. **Privacy and Data Security**
3. **Control and Customization**
4. **Dependence on Connectivity**
5. **Transparency and Explainability**
6. **Ethical and Regulatory Compliance**
7. **Vendor Lock-in and Availability**
8. **Environmental Impact**

$$\hat{q}_t = \arg\max_{q_t} p(q_t|c,t) = \arg\max_{q_t} \sum_{i=1}^{|q_t|} \log p(w_i|\mathbf{c},\mathbf{t},w_1\dots w_{i-1})$$

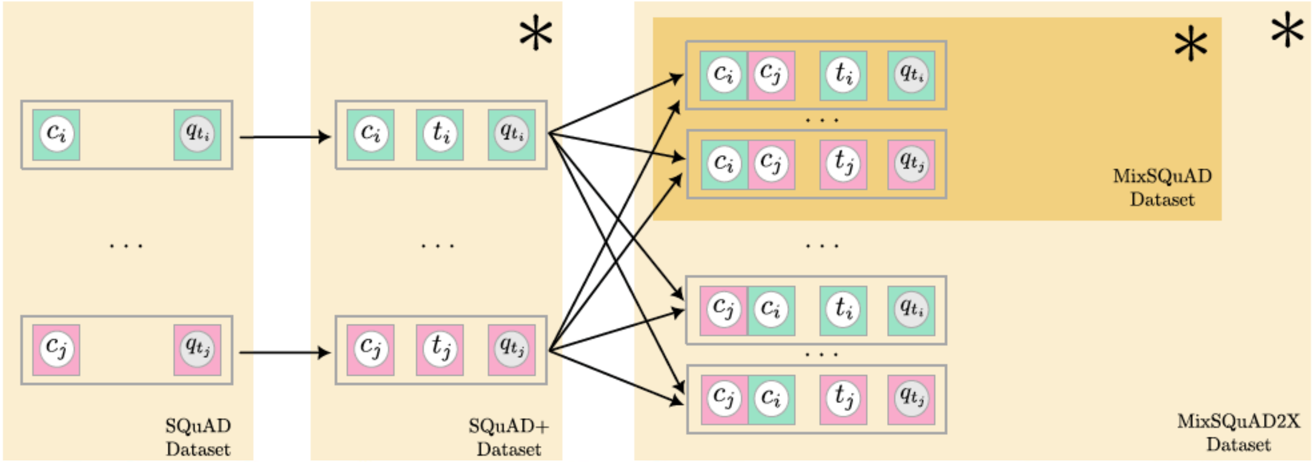


Table 4. Semantic relatedness between the generated questions $\hat{q}$ on (i) prescribed topic t vs. (i) alternative topic t' and the reference question on the prescribed topic q_t. The best performance and the next best for each metric is highlighted in **bold** and *italic*.

	BERTScore			WikiSimRel (Jaccard)		
	$\hat{q}_t \uparrow$	$\hat{q}_{t'} \downarrow$	Difference $\uparrow$	$\hat{q}_t \uparrow$	$\hat{q}_{t'} \downarrow$	Difference $\uparrow$
Baseline	**0.859**	0.859	0.000	0.615	***0.070***	0.545
TopicQGedu	0.855	0.831	0.024	0.721	0.185	0.536
TopicQG	**0.859**	*0.830*	*0.029*	*0.727*	*0.132*	*0.595*
8bit	*0.858*	0.831	*0.027*	0.693	0.142	0.551
4bit	*0.858*	0.831	0.027	0.686	0.157	0.529
TopicQG2X	**0.859**	**0.823**	**0.036**	**0.735**	**0.055**	**0.680**

Li, Z., Cukurova, M., Bulathwela, S. (2025). A Novel Approach to Scalable and Automatic Topic-Controlled Question Generation in Education. *Learning Analytics & Knowledge Conference*, ACM New York.

AI can provide teachers with productivity gains in content generation, but qualitative improvements in practice at a scale are yet to be evidenced.

Student-facing AI: A Learning Sciences-driven Approach to Intelligent Tutoring

- Productivity gains of task completion are of secondary importance to students.

- LLMs should integrate core principles of effective human learn learning.

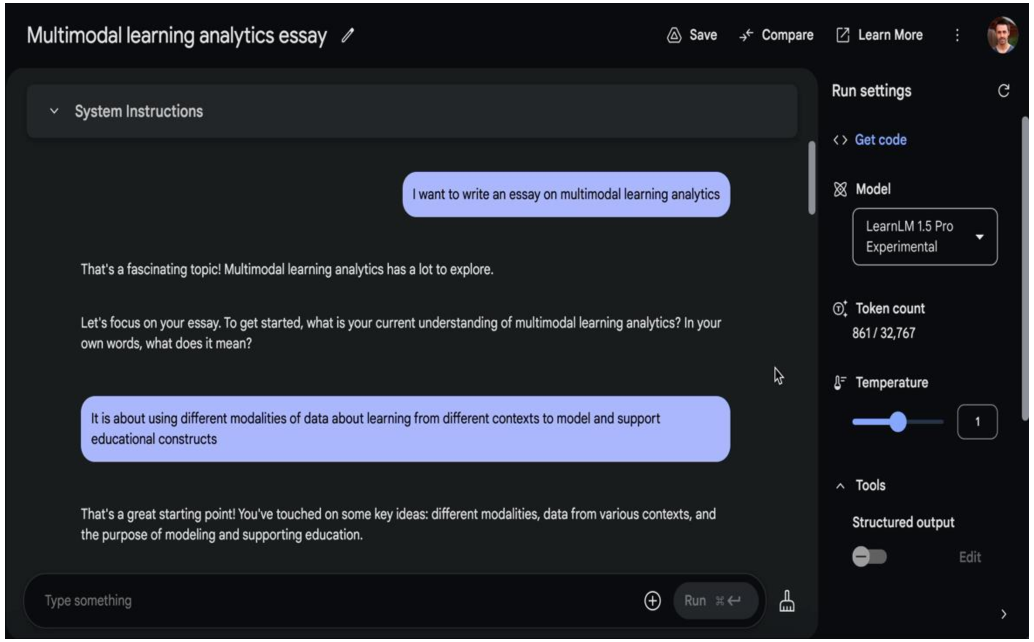
- **Student-facing AI should NOT be "answer machines",** but should provide step-based learning, tracking students' thought processes step-by-step, rather than just checking final answers.

- The importance of targeted and specific feedback, cognitive load, productive failure, spaced practice, interweaving, SRL development etc.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive Tutors: Lessons Learned. *The Journal of the Learning Sciences, 4*(2), 167-207.
Luckin, R., & Cukurova, M. (2019). Designing educational technologies in the age of AI: A learning sciences-driven approach. *British Journal of Educational Technology*, *50*(6), 2824-2838.

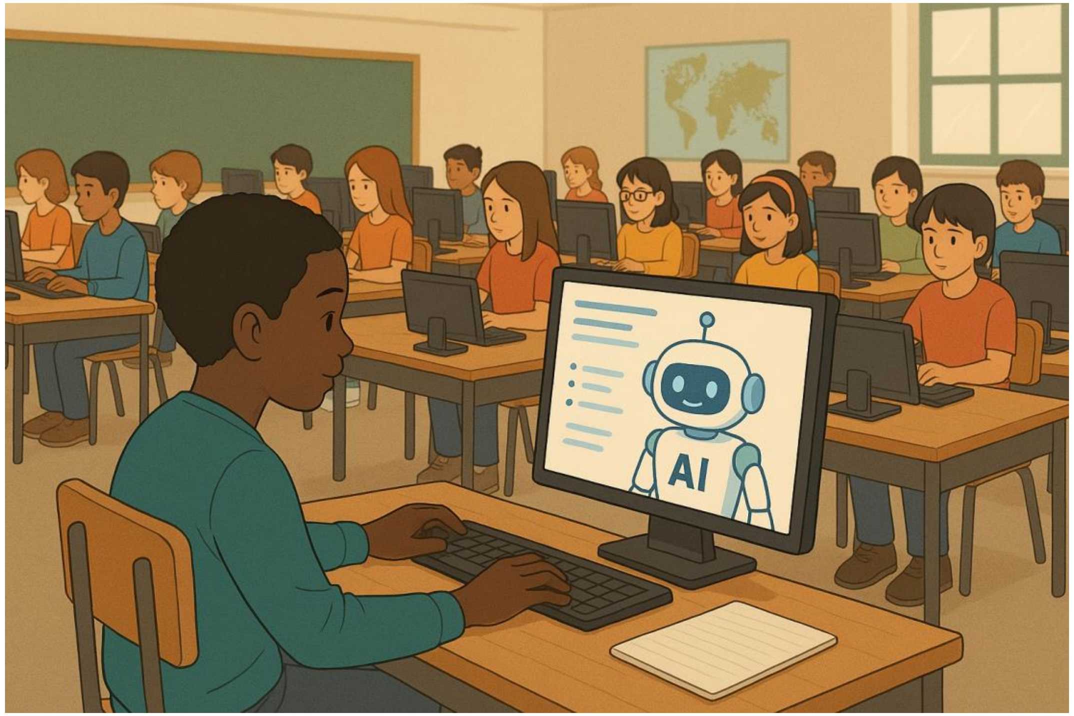
1) Still an Active Research Topic: How do we steer the LLMs beyond system prompts and basic RAG approaches?

- Training a new LLM is not realistic for each pedagogical task (e.g. Google LearnLM trained to align with learning science principles).
- Parameter-Efficient Fine-Tuning (e.g. LoRa, QLoRa) doesn't appear to be enough for complex pedagogical acts.
- Structured knowledge from graphs and ITSs to be embedded in different stages of a GenAI pipeline?
- NeuroSymbolic Approach?



Even if we address them all, what is the future of education?

Learning is not only about knowledge acquisition, and education is not only about learning.



Cukurova, M. (2024). The Interplay of Learning, Analytics, and Artificial Intelligence in Education. *British Journal of Educational Technology*. https://doi.org/10.48550/arXiv.2403.16081

Evidence of Impact: Intelligent Tutoring Systems

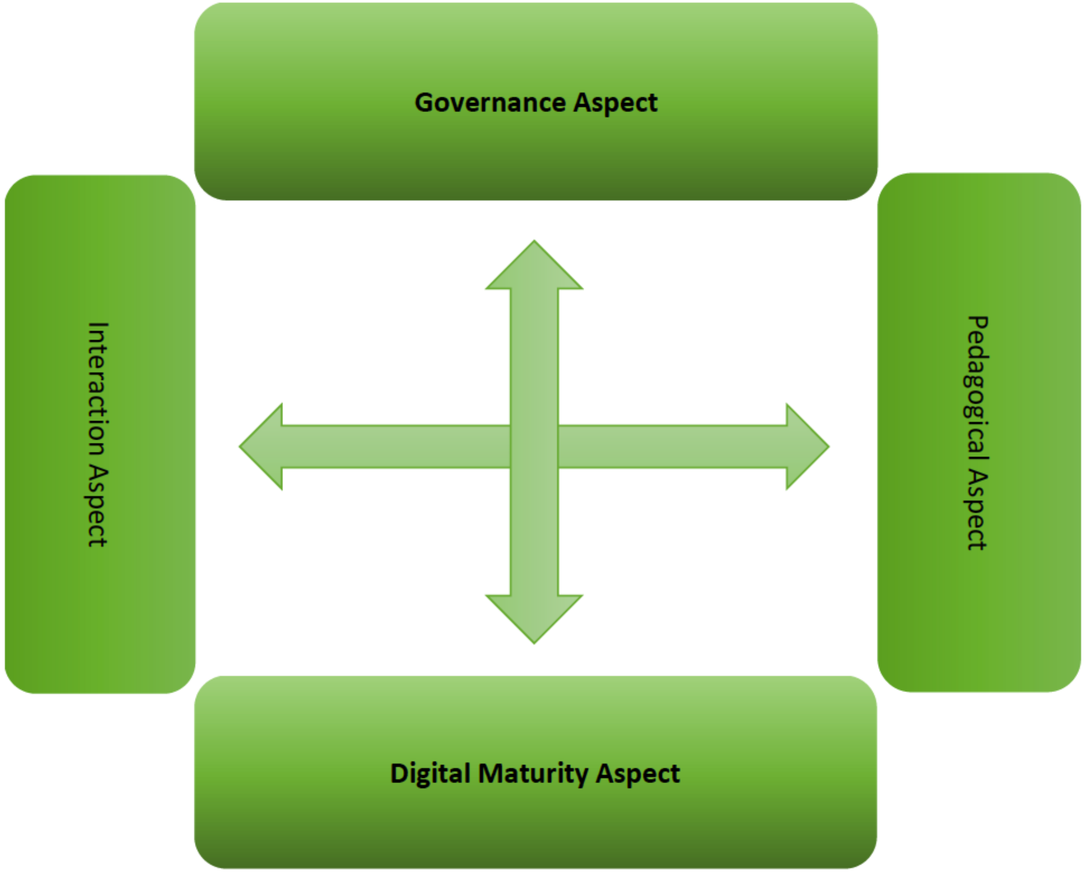
- ITSs can have positive impact on student learning : OLI learning course (Lovett et al., 2008), SQL-Tutor (Mitrovic, & Ohlsson 1999), ALEKS (Craig et al. 2013), Cognitive Tutor (Pane et al. 2014), ASSISTments (Koedinger et al. 2010).

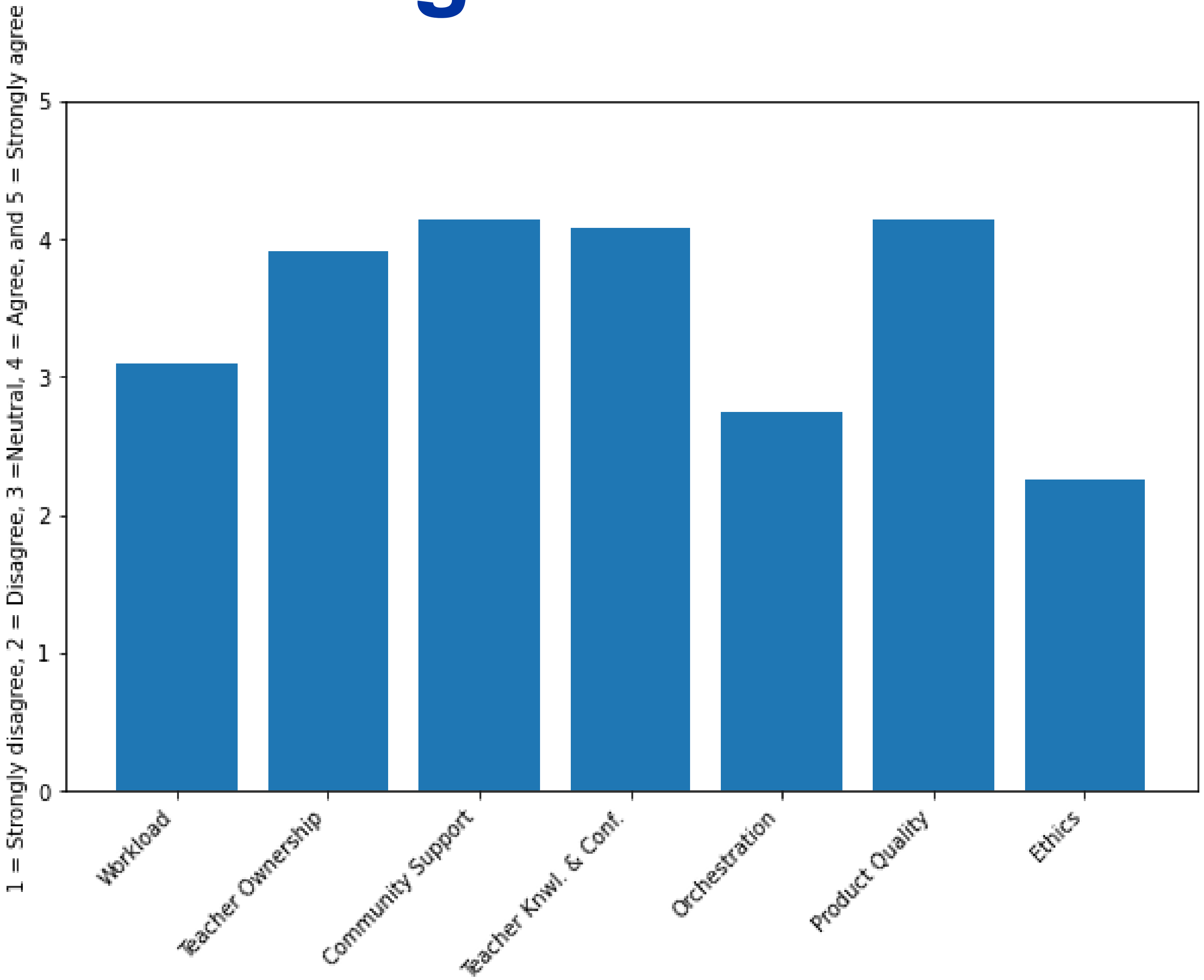
Meta-reviews

- VanLehn (2011) found that the effectiveness of the intelligent tutoring systems were nearly as effective as average human tutors.
- Ma et al. (2014) found similar results both when compared to a no tutoring or to large group human-tutor instruction.
- Pane et al. (2014) found evidence of the relative effectiveness of online tutors over conventional teaching.
- Kulik & Fletcher (2016) median effect was to raise test scores 0.66 standard deviations over conventional levels, or from the 50th to the 75th percentile.
- du Boulay, B. (2016) summary of the metareviews in “Artificial Intelligence As An Effective Classroom Assistant”.

Despite significant advancements in AI and evidence supporting its effectiveness as an ITS, why AI has NOT been prevalent in mainstream education?

AI in education is inherently a socio-technical ecosystem challenge

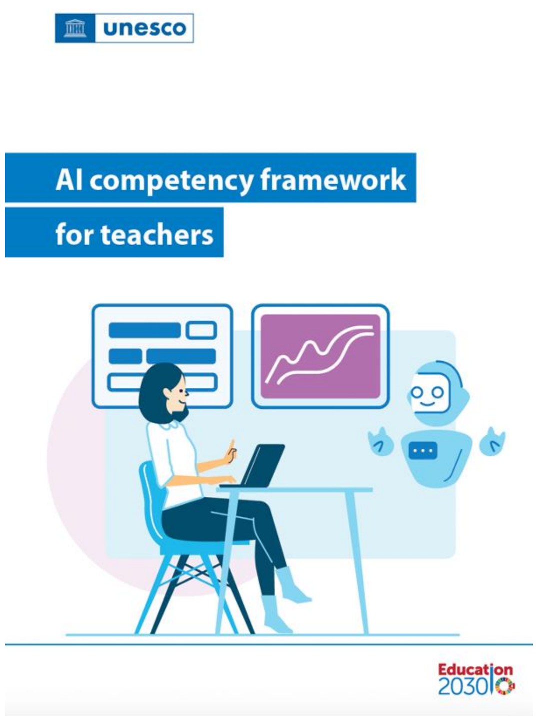




Cukurova, M., Miao, X., & Brooker, R. (2023). Adoption of Adaptive Learning Platforms in Schools: Unveiling Factors Influencing Teachers Engagement. *Artificial Intelligence in Education, Springer.* https://doi.org/10.1007/978-3-031-36272-9_13

AI Competency Framework for Teachers

Aspects	Progression		
	Acquire	**Deepen**	**Create**
Human-centred Mindset	Human agency	Human accountability	AI social responsibility
Ethics of AI	Ethical principles	Safe and responsible use	Co-creating AI ethics
AI Foundations & Applications	Basic AI techniques and applications	Application skills	Creating with AI
AI Pedagogy	AI-assisted teaching	AI-pedagogy integration	AI-enhanced pedagogical transformation
AI for Professional Development	AI enabling lifelong professional learning	AI to enhance organizational learning	AI to support professional transformation



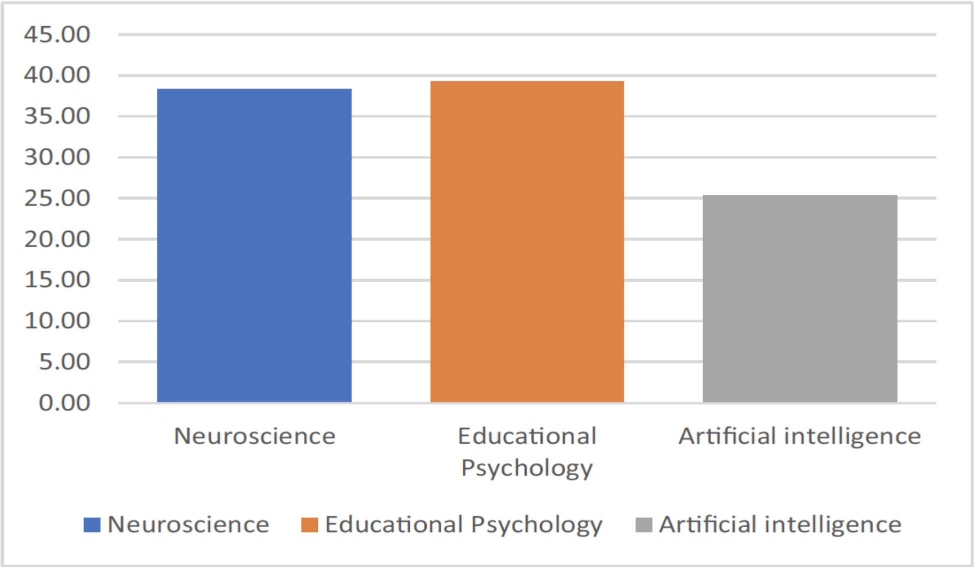
Miao, F. & Cukurova, M. (2024) UNESCO AI Competency Framework for Teachers, UNESCO Publishing

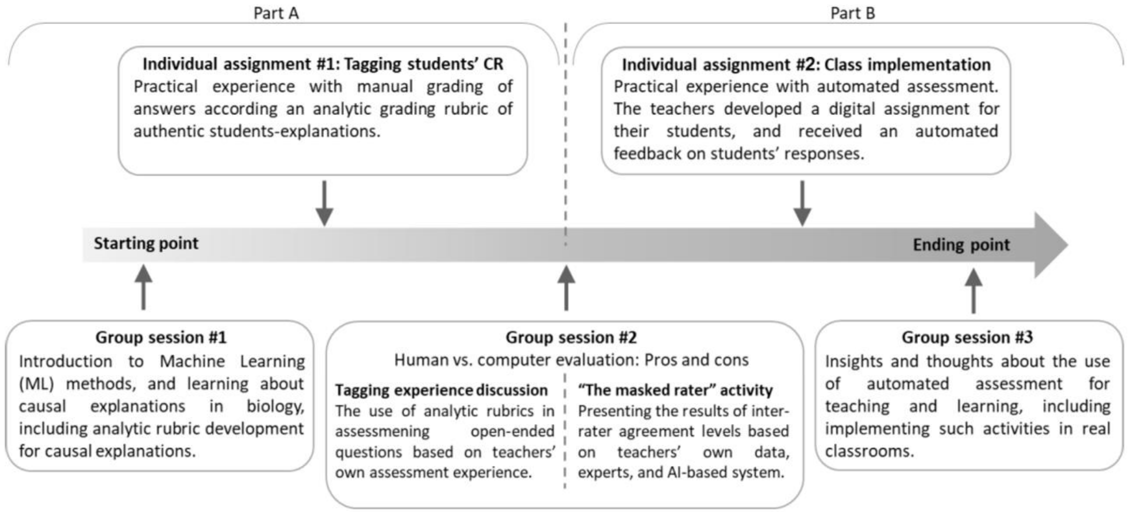
Motivation and Trust Barriers

Students need to be motivated enough to engage with AI tools in the first place, yet only about 5% of them manage to engage with educational content long enough to get statistically significant benefits.

Teachers and learners still have confirmation biases and unrealistic expectations from AI-EdTech.

"AI framing effect": when people are presented with content framed as coming from AI, they tend to judge it as less credible compared to educational psychology and neuroscience.

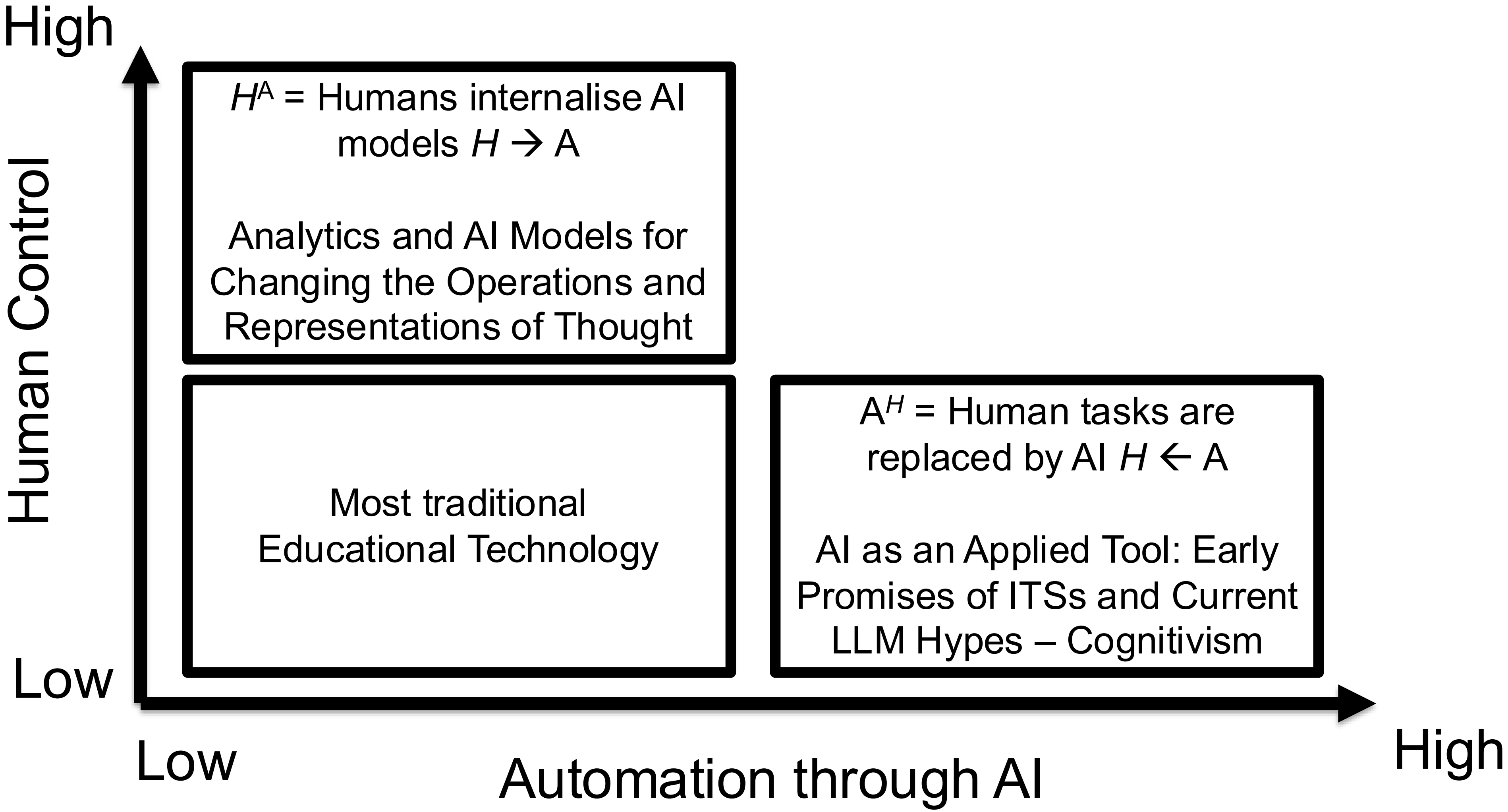




Cukurova, M., Luckin, R., & Kent, C. (2020). Impact of an Artificial Intelligence Research Frame on the Perceived Credibility of Educational Research Evidence. *International Journal of Artificial Intelligence in Education*, 1-31.
Nazaretsky, T., Ariely, M., Cukurova, M., Alexandron, G. (2022). Teachers' Trust in AI-powered Educational Technology and a Professional Development Program to Improve It, *British Journal of Educational Technology*, DOI: 10.1111/bjet.13232
Nazaretsky, T., Cukurova, M., Ariely, M., & Alexandron, G. (2021). Confirmation bias and trust: human factors that influence teachers' attitudes towards AI-based educational technology. In EC-TEL -CEUR Workshop Proceedings (Vol. 3042).

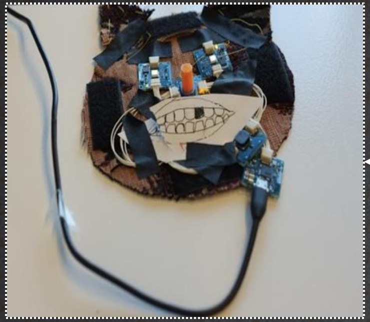
There appears to be limited work in AI in Education focusing on innovative socio-constructivist pedagogies.

AI in Education: A vision for the future

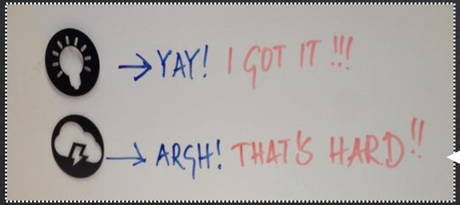


Cukurova, M. (2024). The Interplay of Learning, Analytics, and Artificial Intelligence in Education. *British Journal of Educational Technology*. https://doi.org/10.48550/arXiv.2403.16081

UCL




Available Scenarios
Name
1. Your plan
2. Document your work
3. Reflect


→ YAY! I GOT IT !!!
→ ARGH! THAT'S HARD!!

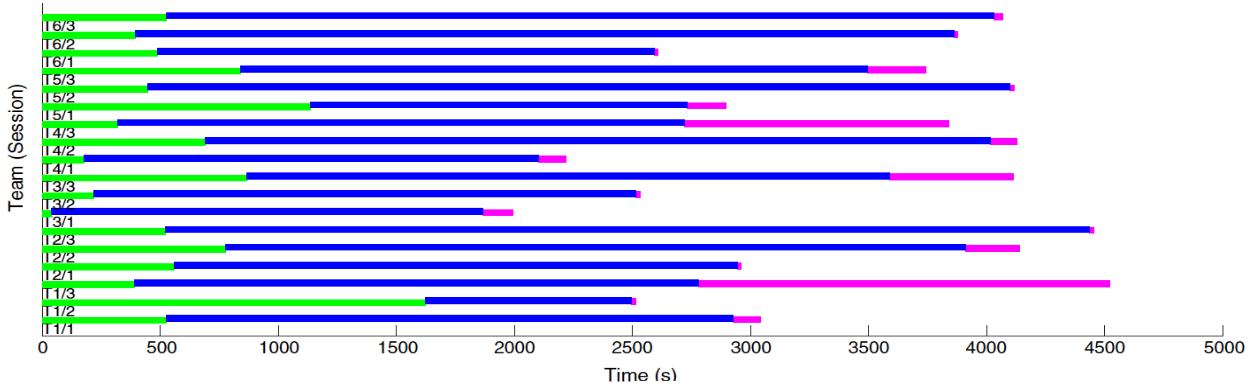
It may never be possible to build a model to accurately predict how learning in a complex social context evolves.

Independent Variables (MMLA Features)

FLS - Number of faces looking at screen
DBL - Mean distance between learners
DBH - Mean distance between hands
HMS - Mean hand movement speed
AUD - Mean audio level

IDEX - Arduino measure of complexity
IDEVHW - Arduino active hardware blocks
IDEVSW - Arduino active software blocks
IDEC - Arduino active blocks

PWR - Student Work Phases



Method	Deep learning	Traditional
Task	Regression	Classification
Input	18 variables	9 variables per window
Output	6 scores over 5 levels	1 score with 3 levels
Metrics	Regression score	Classifier accuracy
Windowing	120,240 and 360 s	10,20,30,90 min
Phase exclusion	Reflection	Reflection
Method	Multiple layers	NB, LR, SVML, and SVMR

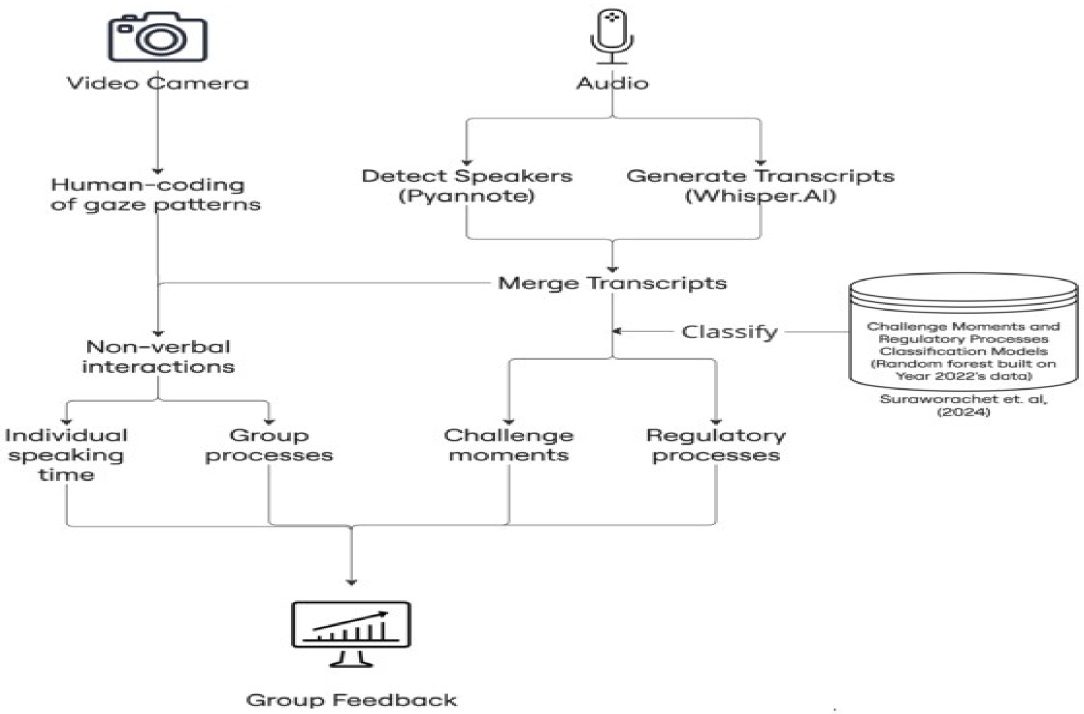
Note. NB = naive Bayesian; LR = logistic regression; SVML = support vector machines with linear kernel; SVMR = support vector machines for regression.

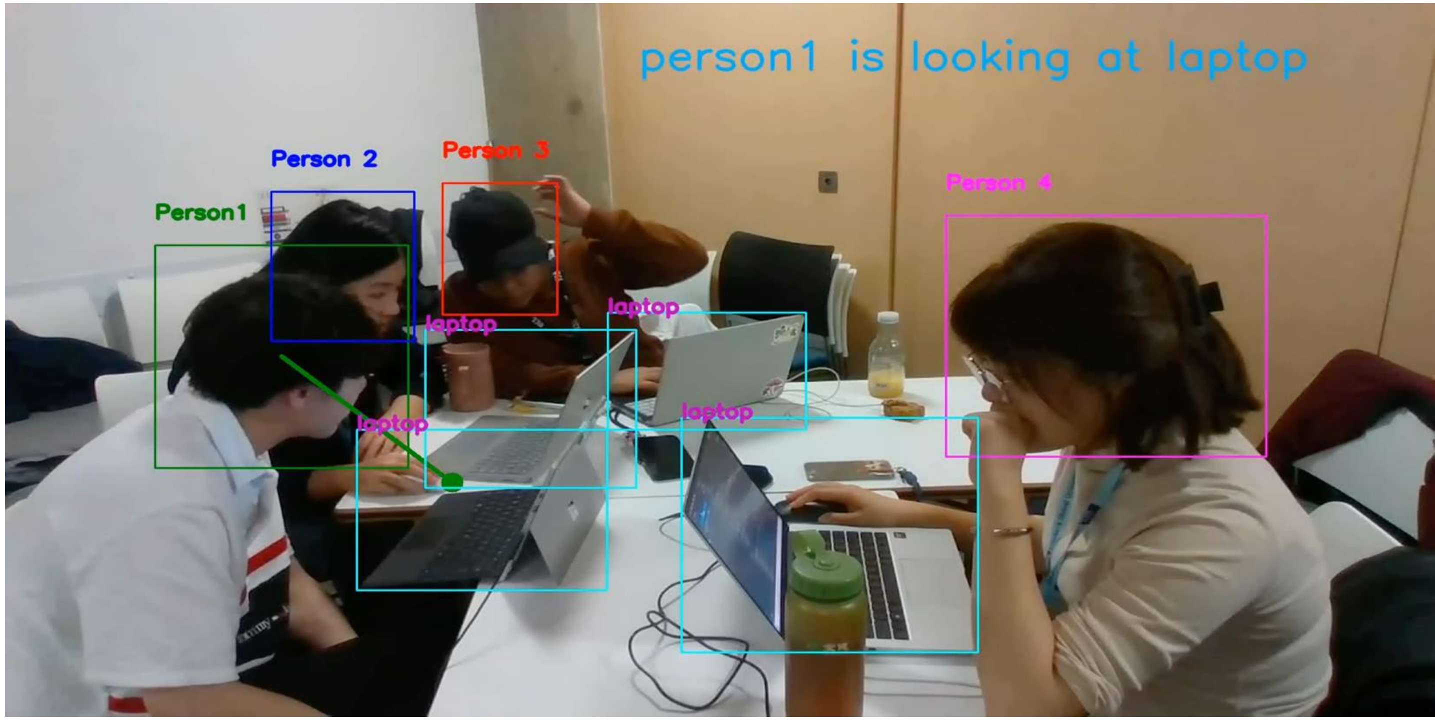
Removed feature	Best result
No features removed	0.129
All faces data	0.21
All Arduino data	0.21
DBF	0.15
DBH	0.21
HMS	0.19
AUD	0.18
Hand pos	0.21
Arduino comp	0.19

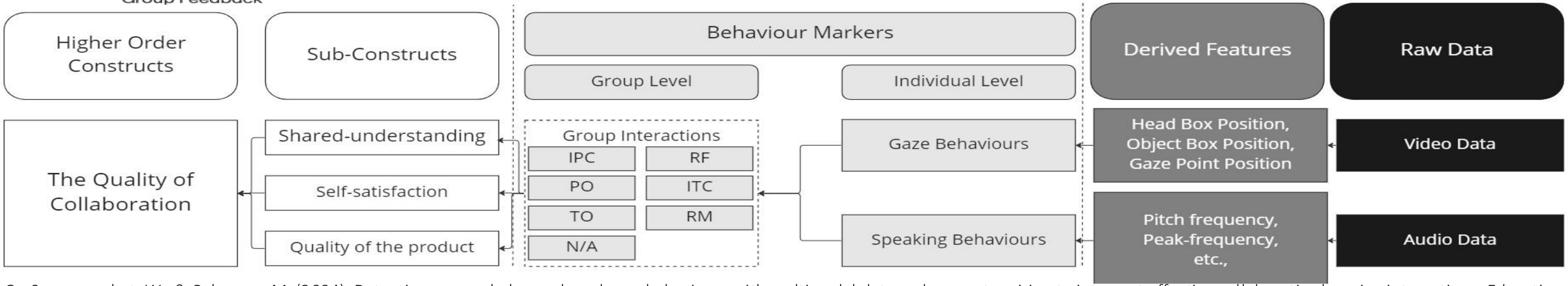
Ground Truth: Expert labelling of video data using CPS frameworks

Spikol, D., Ruffaldi, E., Dabisias, G., & Cukurova, M. (2018). Supervised machine learning in multimodal learning analytics for estimating success in project-based learning. *Journal of Computer Assisted Learning*, *34*(4), 366-377.

Decompositionality principle and Making learning visible





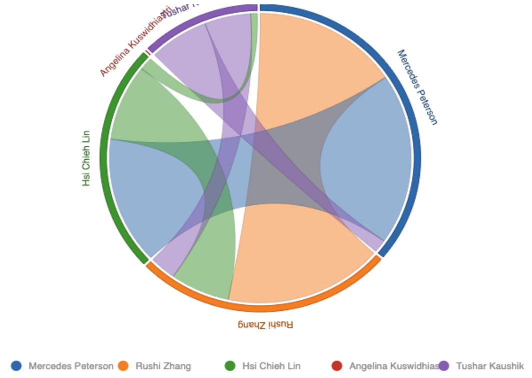


Zhou, Q., Suraworachet, W., & Cukurova, M. (2024). Detecting non-verbal speech and gaze behaviours with multimodal data and computer vision to interpret effective collaborative learning interactions. *Education and Information Technologies*, *29*(1), 1071-1098.

Value of Making Lived Experiences Visible: Students' group interactions and regulation challenges

How did we act?

Mutual discussion (times)

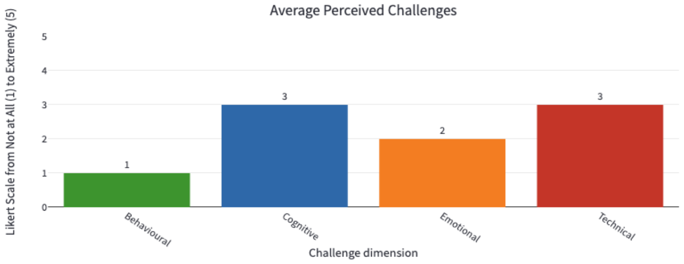


Which group interactions were your group in?

How did we regulate?

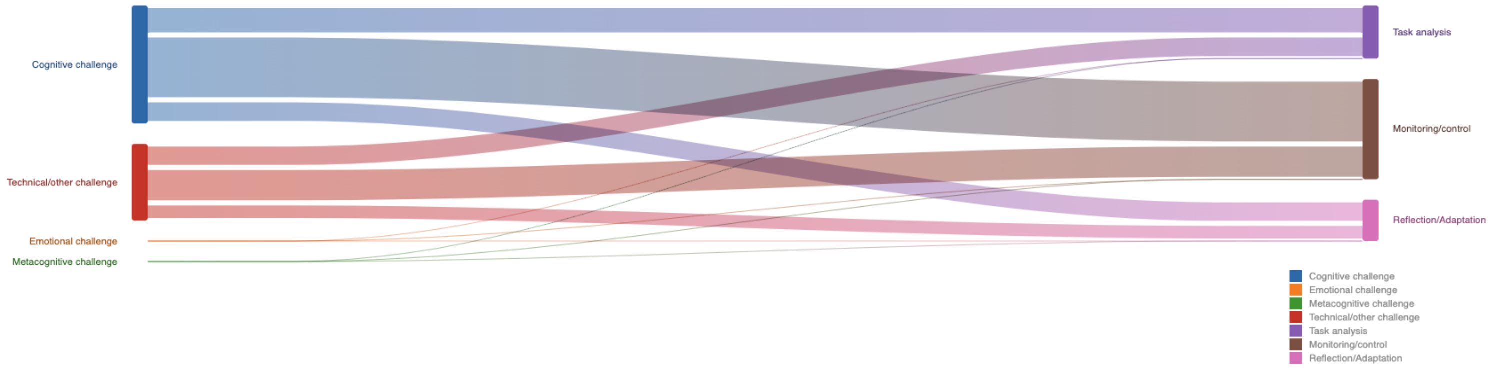
Reported challenges

from post-surveys



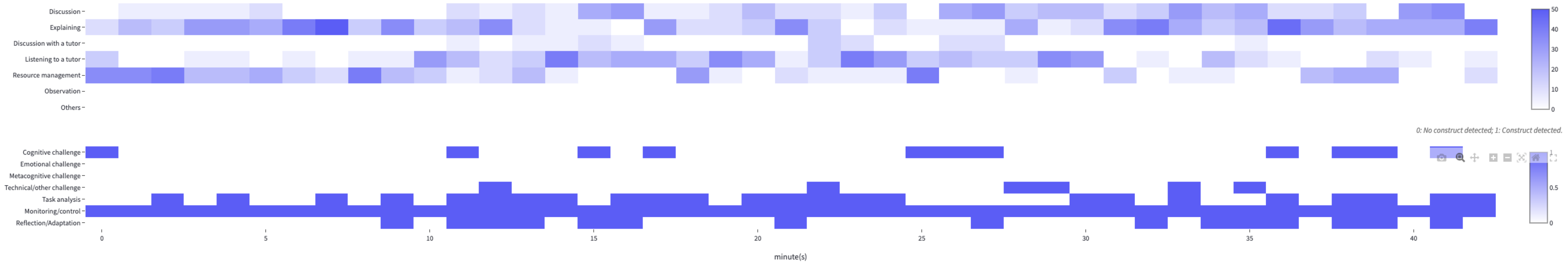
Detected challenges and regulation

from transcript



Suraworachet, W., Seon, J., & Cukurova, M. (2024). Predicting challenge moments from students' discourse: A comparison of GPT-4 to two traditional natural language processing approaches. *Learning Analytics & Knowledge, ACM: New York*.

Feedback Generation on Observed States



Post-survey

- This week, the post-survey response rate reached **20%**. A response rate closer to 100% offers a stronger representation of the group's collective perception. Overall, the group's satisfaction level with the tasks and products is satisfied. Continued engagement helps ensure all voices are well-represented in shaping the collaborative experience!

Mutual discussion

- The chord diagram illustrates the dynamics of mutual discussions within the group, highlighting the frequency with which members engaged by responding to or following up on one another's contributions. The most active engagement occurred between **Mercedes Peterson-Hsi Chieh Lin**, while **Tushar Kaushik-Hsi Chieh Lin** showed less interaction. This highlights the importance of balanced and consistent engagement among members to promote positive group collaboration.

Group interactions

- The pie chart shows the proportion of group interactions detected during the session.
- According to the pie chart, your group invested a significant amount of time in both listening to each other explain relevant concepts based on the learning materials and engaging in discussions. It's great to observe that these discussions and explanations occurred in turns, indicating that group members were actively contributing to each other's points of view to negotiate meaning and work toward building a shared understanding.
- The graph indicates that your group experienced some periods with no collaboration. This is perfectly fine and may be attributed to many reasons including separate discussions occurring during the process. According to previous literature, successful collaboration is built on the consensus of each member through inclusive discussions. It would be helpful to aim for including all members of your group in the discussion.

Detected challenges/regulation

- The Sankey chart illustrates the transition from identified challenges to corresponding regulatory processes.
- Notably, there were significant transitions between **Cognitive challenge-Monitoring/control.**
- While specific recommendations linking challenges to regulatory processes are not provided, it's beneficial to consider the context of each challenge. This can help in tailoring self-regulation, peer support, or group adjustments to respond effectively and adapt to the challenges encountered.

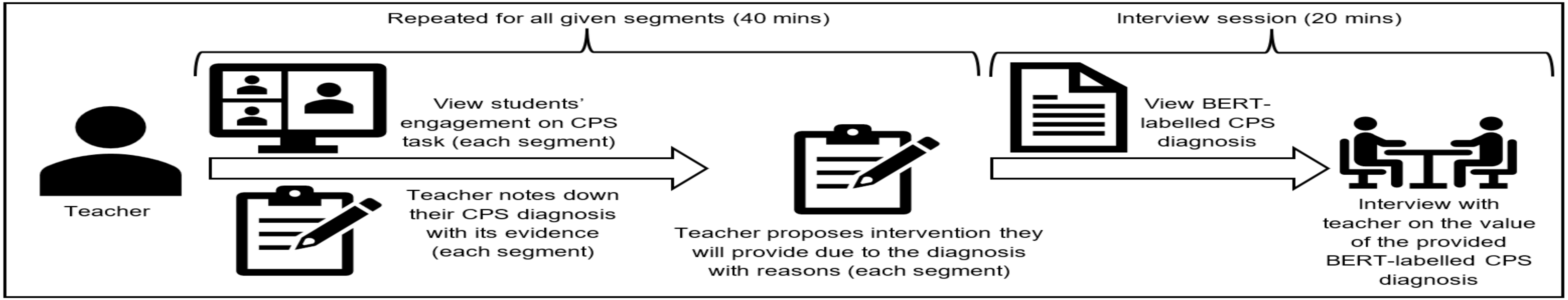
Uzun, Y., Zhou, Q., Suraworachet, W., Gauthier, A., & Cukurova, M. (2025). Engagement with analytics feedback and its relationship to self-regulated learning competence and course performance. *Int J Educ Technol High Educ* , *29*(1), 1071-1098.

Social-Cognitive Dimension	Subskill	Label	Indicator
Problem Solving Subskill	SS1: Sense-making	PS01	Talking about the task questions in general terms to understand about the problem-solving task
		PS02	Explaining ideas or concepts in the problem-solving task with reference to prior knowledge or definitions from information sources
		PS03	Addressing difficulties or limitations that obstruct problem solving
	SS2: Building shared understanding	PS04	Asking questions to clarify understanding, ideas or contributions
		PS05	Answering questions to clarify understanding, ideas or contributions
		PS06	Reiterating or paraphrasing oneself or others' ideas or contributions
		PS07	Adapting and building on the ideas or contributions of others
		PS08	Stating agreement with others
		PS09	Discovering perspectives and abilities of team members
		PS10	Sharing information from sources which contributes to formulating the problem-solving task
		PS11	Stating disagreement with others
		PS12	Constructing arguments in favour of one's own ideas or contributions
		PS13	Resolving differences
		PS14	Reaching a compromise with others
		PS15	Identifying and abstracting relevant information about the task context
		PS16	Establishing connections and patterns between relevant information in the problem-solving task
		PS17	Dissecting the problem into smaller tasks
	SS3: Formulating a solution	PS18	Building a representation of the problem-solving task
		PS19	Creating an ordered step-by-step plan
		PS20	Proposing ideas or specific solutions methods to solve the task questions
	SS4: Defining roles and responsibilities	PS21	Discussing required roles and collaborative interaction to address the problem-solving task
		PS22	Coordinating sub-tasks to be performed
	SS5: Reaching a solution	PS23	Sharing contributions and findings of individual and team sub-tasks
		PS24	Providing an answer to the task questions
		PS25	Responding to or acknowledging the contributions of others
	SS6: Maintaining roles and responsibilities	PS26	Discussing the progress and status of individual and team sub-tasks
		PS27	Providing feedback on the progress and status of individual or team sub-tasks
		PS28	Recognising strengths and weaknesses of self and others
		PS29	Adapting team organisation to adjust individual and team sub-tasks
	SS7: Maintaining shared understanding	PS30	Providing feedback or instructional support to others
		PS31	Using feedback provided to clarify or elaborate own ideas
		PS32	Making iterative adaptations to the plan based on outcomes, new information and new ideas
	SS8: Evaluating the solution	PS33	Anticipating issues or errors
		PS34	Testing to detect working order
		PS35	Detecting and hypothesising issues or errors
		PS36	Identifying the need for additional information, resources or tasks to address issues or fix errors

Dimension	Social-Cognitive				Affective			
Model	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
RF TF-IDF	.524	.514	.524	.468	.866	.851	.866	.852
RF TF-IDF+A	.512	.459	.512	.455	.853	.831	.853	.842
BERT	*.589*	*.573*	*.589*	*.573*	**.892**	**.894**	**.892**	*.887*
AudiBERT	**.598**	**.587**	**.598**	**.587**	*.889*	*.890*	*.889*	**.889**

Table 2. Comparison of predictive Accuracy (Acc.), Precision (Prec.), Recall (Rec.) and F1-Score (F1) between models on testing data. The best and second best performances are indicated in **bold** and *italic* faces respectively.

- Compare the sequences of labelled CPS actions for successful groups and present the diagnosis of these patterns to teachers.



Wong, K., Wu, B., Bulathwela, S. & Cukurova, M. (2025). Rethinking the Potential of Multimodality in Collaborative Problem Solving Diagnosis with Large Language Models. *International Conference of Artificial Intelligence in Education*, Springer, Cham.

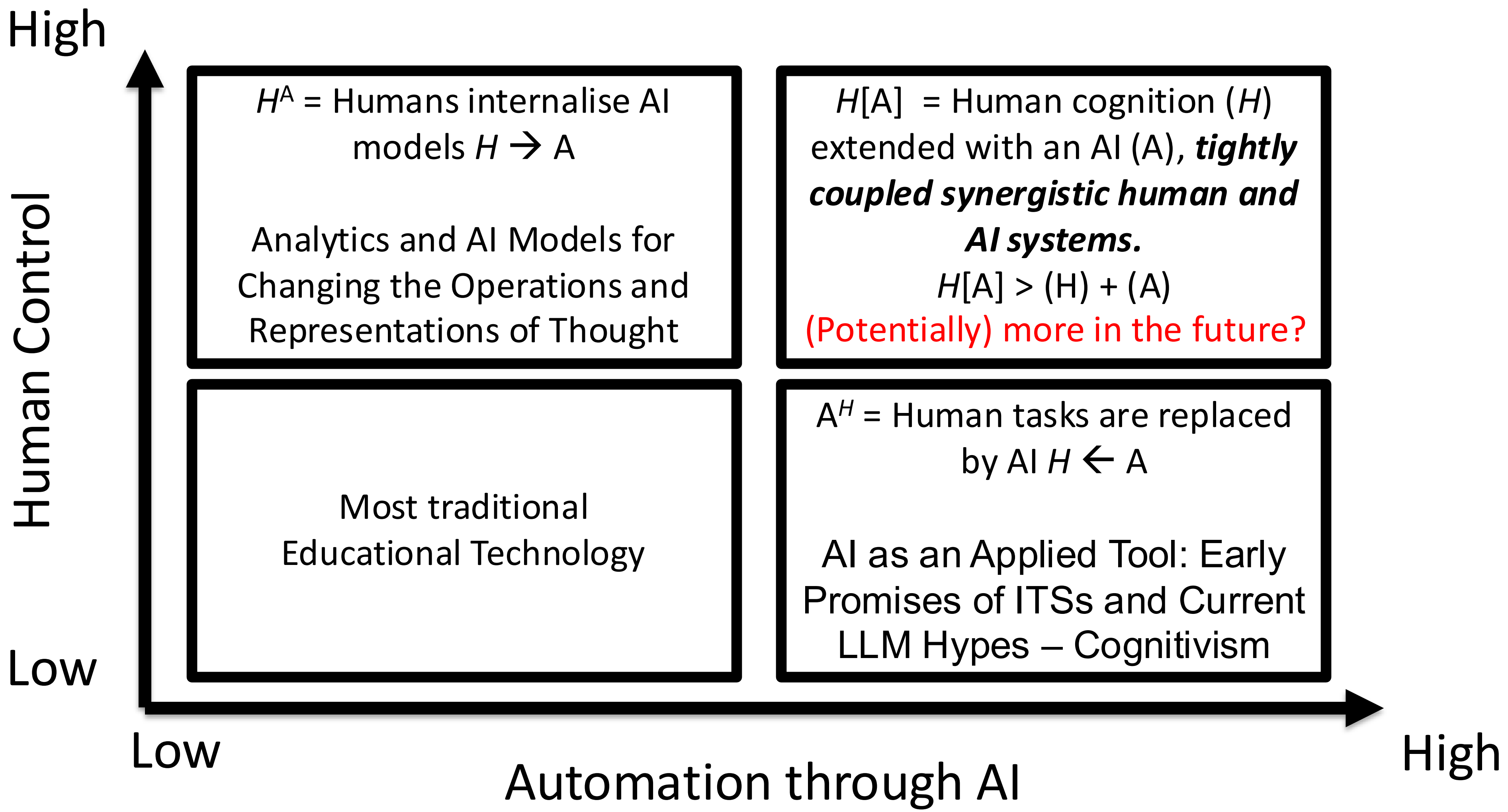
AI models can also help us describe learning behaviours & processes more precisely to make the lived experiences more visible.

The Social Translucence Theory

Visibility	•**Comprehensibility** of the collaboration analytics *(easy to understand/interpret)* •**Accuracy/Inaccuracy** of the analytics information *('Similar to their findings', different from lived experiences)* •**Lack of qualitative feedback** and partially represented contribution *(contribution is more than observed, speak more doesn't mean more contribution)*
Awareness	•The value of **seeing one's own performance** *(as external reflective tool that cannot be distorted by observers/post-experienced effects)* •The value of **seeing others' performance** *(determine who's struggling)*
Accountability	•Collaboration analytics to **foster group discussions** *(discuss why contribute less)* •**Self-regulation** *(adjust level/prepare more/seek for help)* and **socially shared regulation** of behaviours *(encourage the least speaker, offer helps, develop group strategies e.g. host)* •**Gaming the system** *(particularly for speech time data – is it bad?)* •**Swinging back** to "normal" behaviours *(lack of monitoring/assessment)*

Zhou,Q., Suraworachet, W., Pozdniakov, S., Martinez-Maldonado, R., Bartindale, T., Chen, P. Richardson, D., & Cukurova M. (2021). Investigating Students' Experiences with Collaboration Analytics for Remote Group Meetings. *International Conference of Artificial Intelligence in Education*, Springer, Cham.
Pozdniakov, S., Martinez-Maldonado, R., Shan-Tsai, Y., Cukurova, M., Bartindale, T., Chen, P., Harrison, M., Richardson, D., & Gasevic, D. (2022). The Question-driven Dashboard: How Can We Design Analytics Interfaces Aligned to Teachers' Inquiry?. *Learning Analytics & Knowledge, ACM*.

AI in Education: A vision for the future

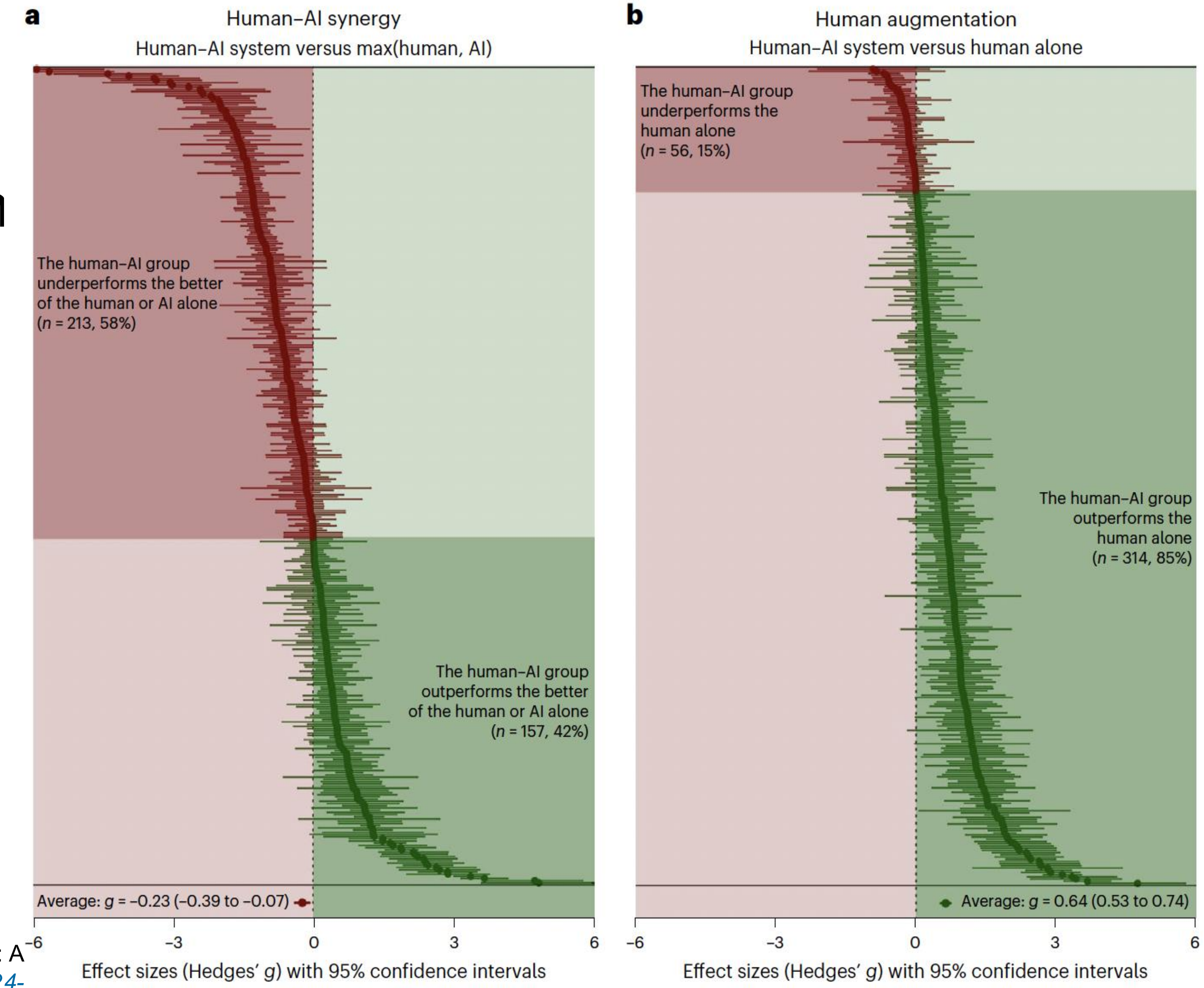


Cukurova, M. (2024). The Interplay of Learning, Analytics, and Artificial Intelligence in Education. *British Journal of Educational Technology*. https://doi.org/10.48550/arXiv.2403.16081

Not all human-AI teaming is synergistic

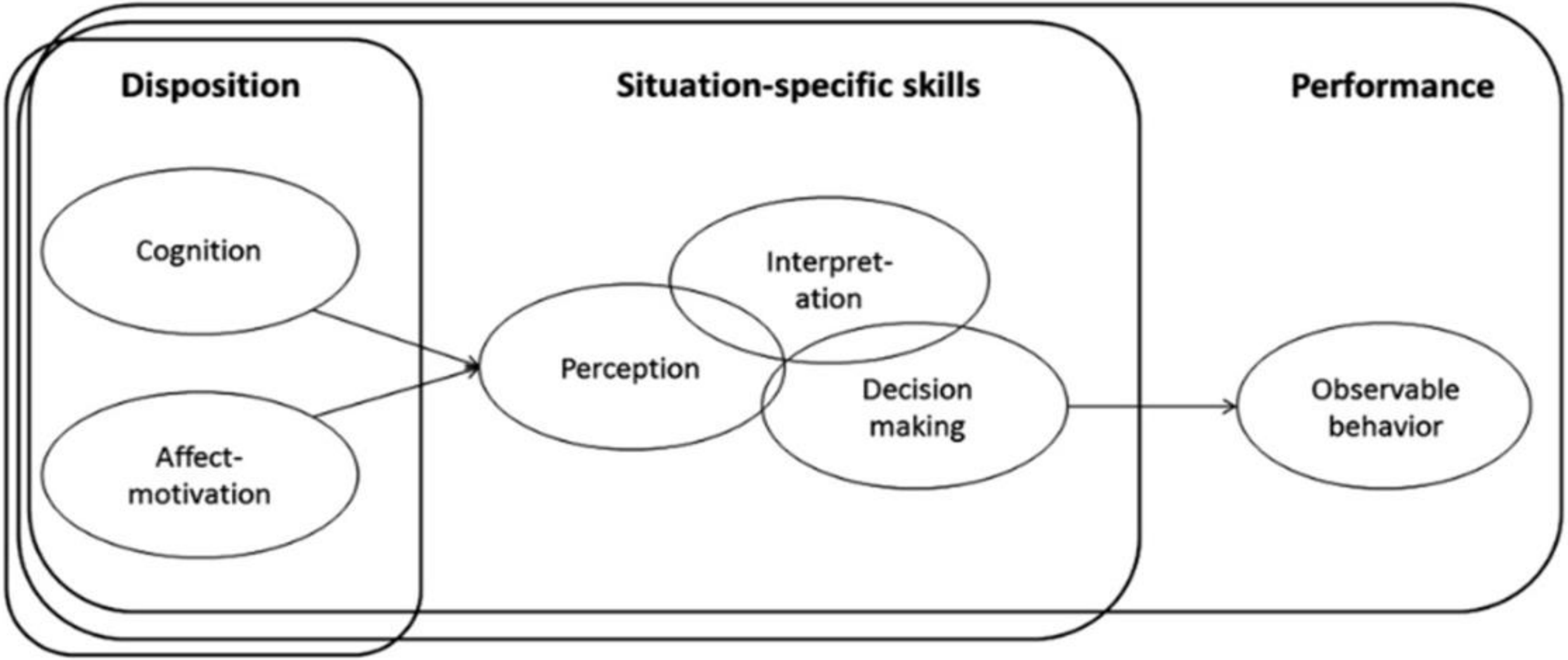
- On average, human–AI teaming performs **significantly worse** than the best of humans or AI alone; significantly better than the human alone (**human augmentation**).

- Performance losses in tasks that involved **making decisions** and significantly greater gains in tasks that involved **creating content**.



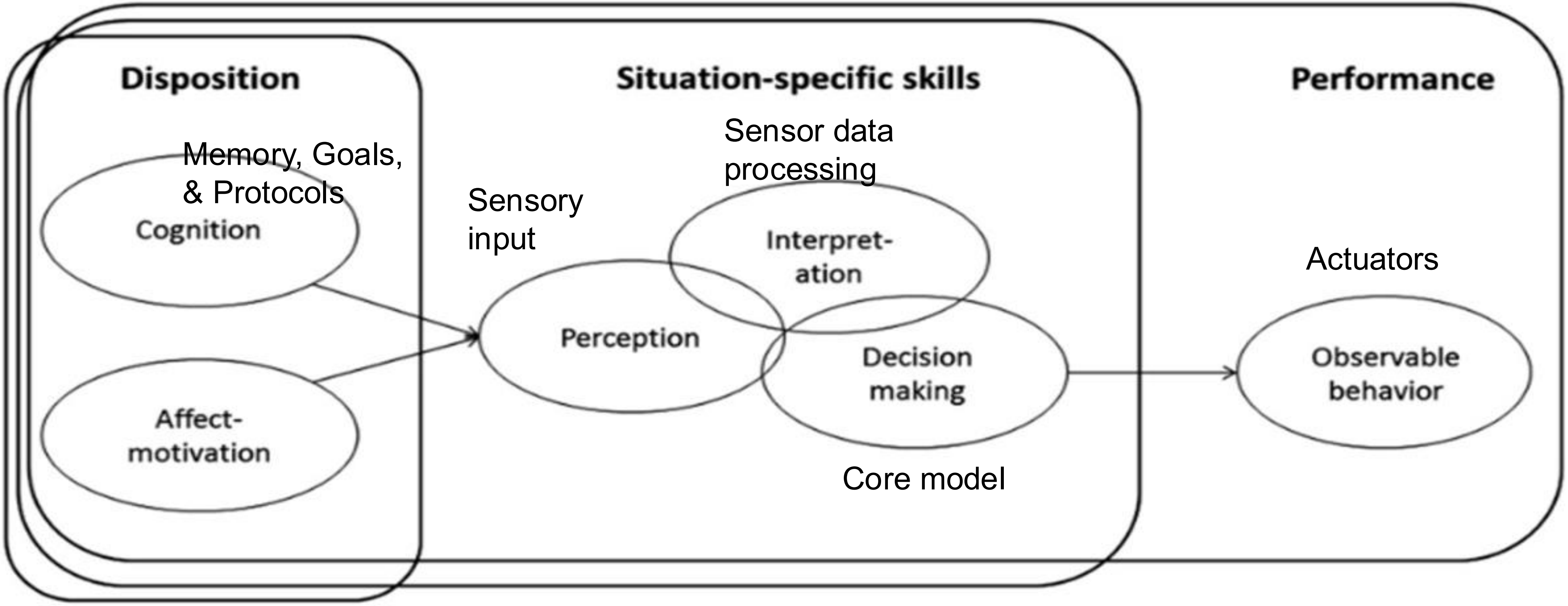
Vaccaro, M., Almaatouq, A., & Malone, T. (2024) When combinations of humans and AI are useful: A systematic review and meta-analysis, *Nature Human Behaviour*, https://doi.org/10.1038/s41562-024-02024-1

Not all human-AI teaming is synergistic



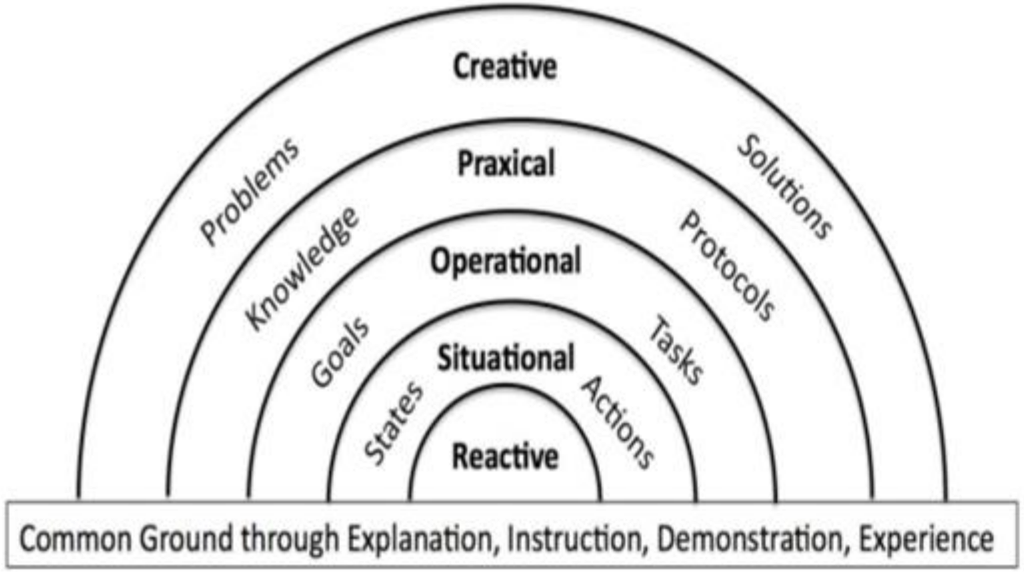
Blömeke et al., (2015) Competence as a continuum

Competence augmentation with AI



In a given task, depending on the specific human competence and AI's affordances;

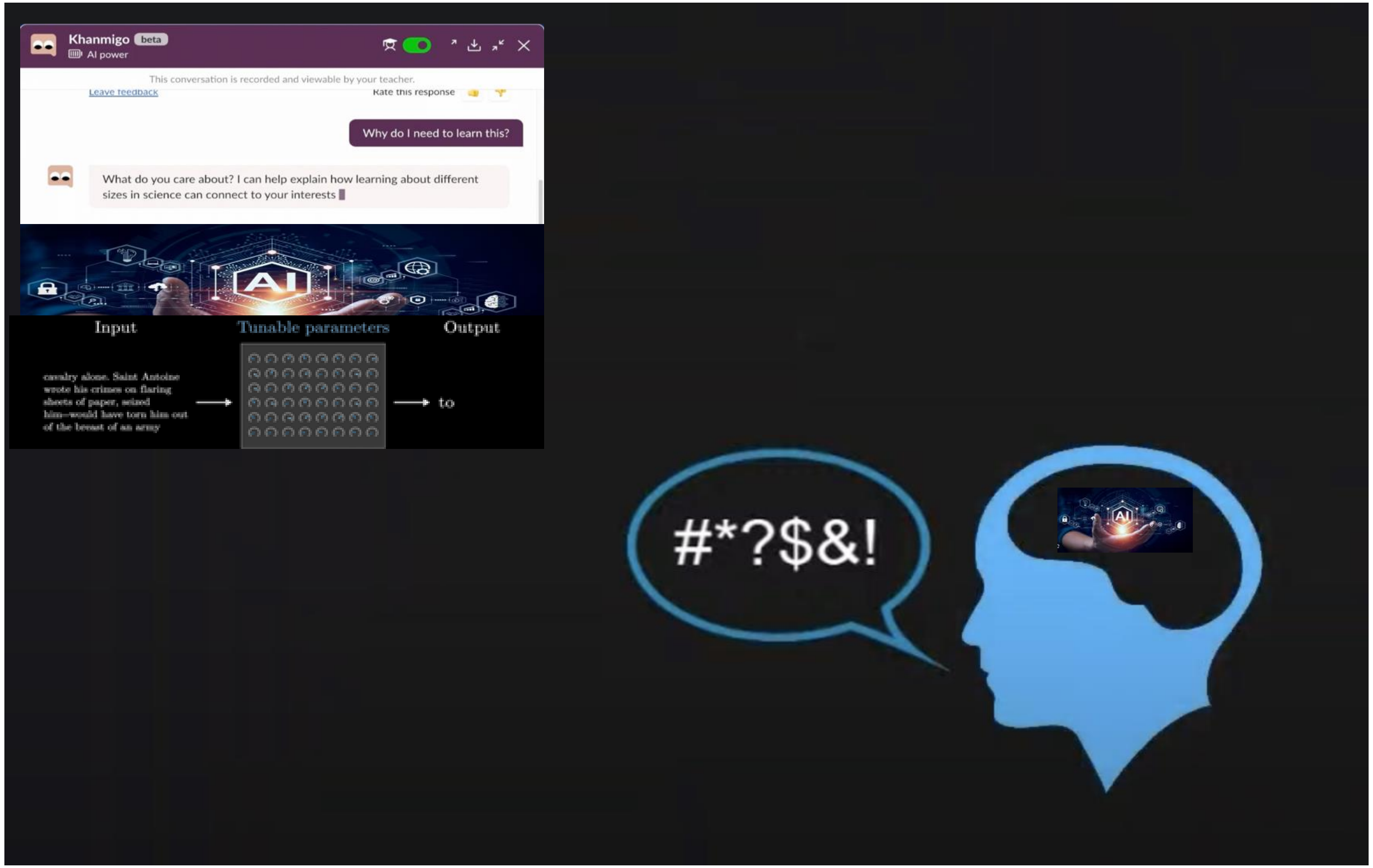
1. Reactive Teaming
2. Situational Teaming
3. Operational Teaming
4. Praxical Teaming
5. Synergistic Collaboration

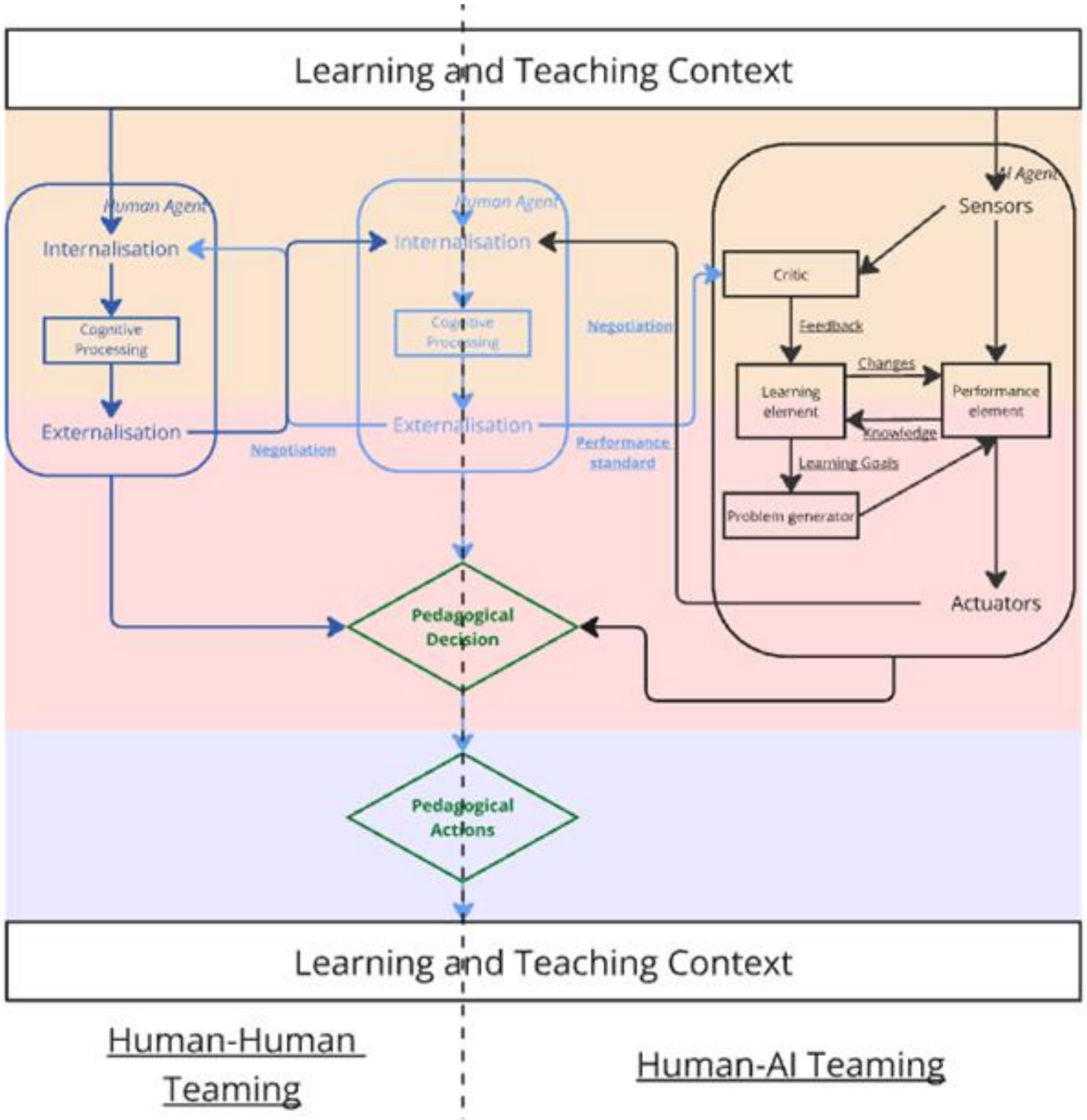


Crowley, J. L., Coutaz, J., Grosinger, J., Vazquez-Salceda, J., Angulo, C., Sanfeliu, A. & Cohn, A. G. (2022). A hierarchical framework for collaborative artificial intelligence. IEEE pervasive computing, 22(1), 9-18.

AI lacks a theory of mind to interact with humans as humans interact with other humans.

Maybe we all need a "theory of AI"?





Synergistic Collaboration with Intelligent Agents

Step 1: Observe: The agent observes the current state of its environment via sensors.

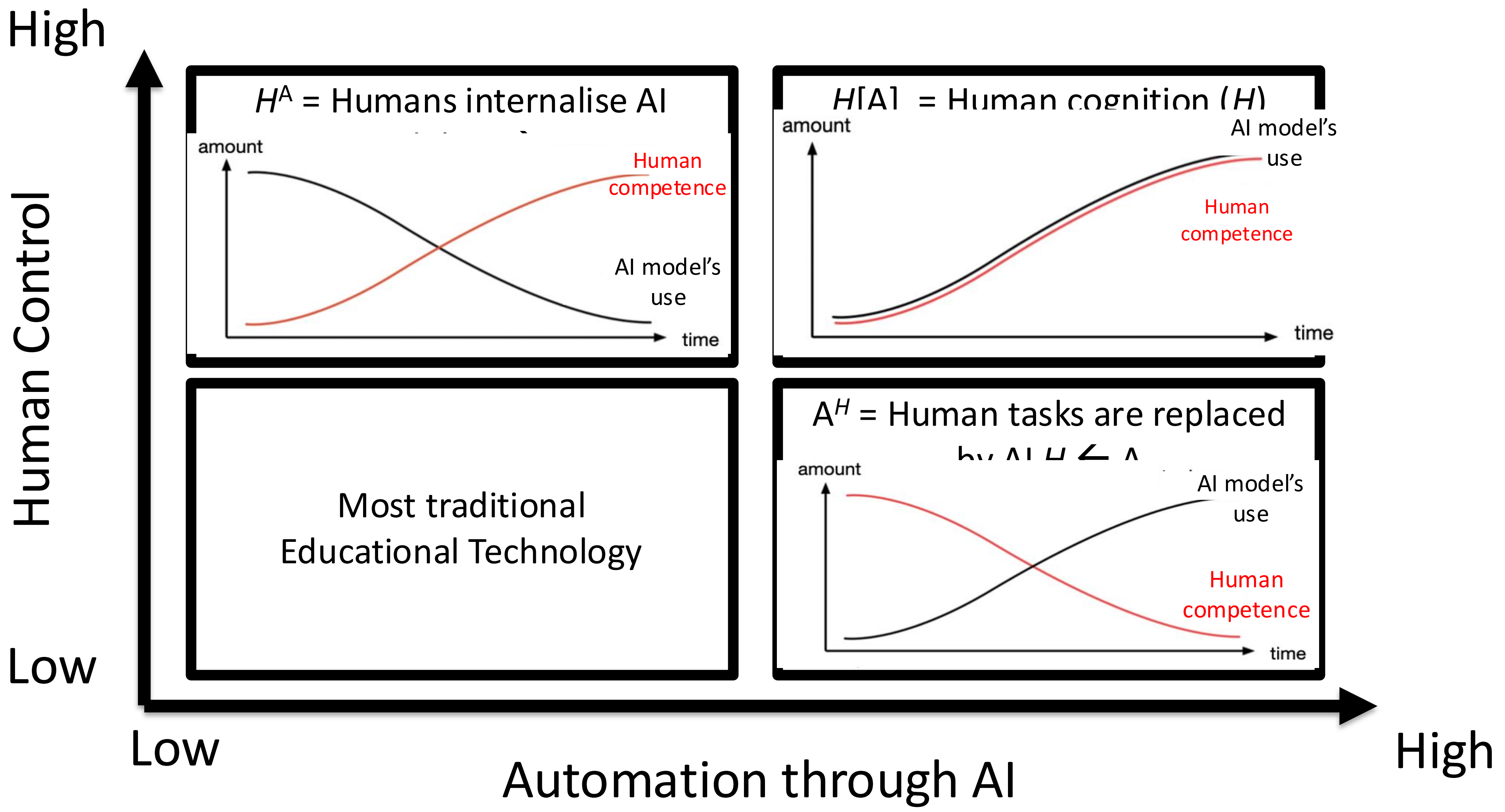
Step 2: Act Suggestions: Based on its current policy or model, the agent selects and proposes an action.

Step 3: Receive Feedback and Negotiate: The agent receives a reward signal, feedback from the human agent and human takes the suggestion on board to negotiate the quality/relevance of AI's chosen action.

Step 4: Mutual Learning and Shared Understanding (Update Model/Policy): Using feedback from each other and observed outcomes, the human and AI **update their internal models**. This adjustment aims to maximise a "shared understanding" and future performance of both human and AI agents.

Step 5: Iterate: The agents repeat these steps continuously, progressively refining their ability to make better decisions.

AI in Education: A vision for the future



Cukurova, M. (2024). The Interplay of Learning, Analytics, and Artificial Intelligence in Education. *British Journal of Educational Technology*. https://doi.org/10.48550/arXiv.2403.16081

CSEDU 2025

17TH INTERNATIONAL CONFERENCE ON COMPUTER SUPPORTED EDUCATION

Thank you

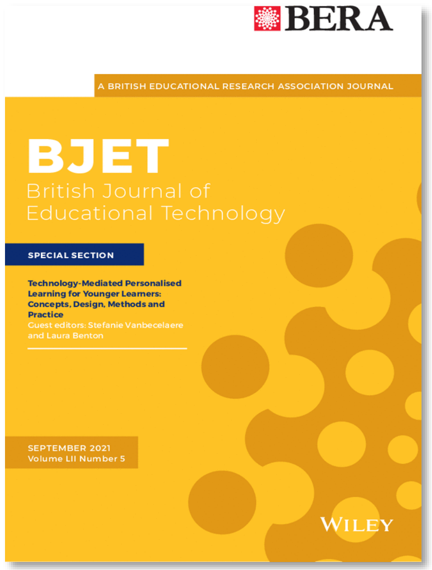
Taico

Teacher-AI Complementarity

Professor Mutlu Cukurova

University College London

m.cukurova@ucl.ac.uk



Knowledge Lab UCL

UCLAIT - Team