Learning and Regulating with ChatGPT: What Experimental Study Tells Us?

Yizhou Fan
Peking University

This study is funded by 2023 SoLAR ECR Grant and Peking University, project: Measuring and Scaffolding Hybrid Human-AI Regulation: Comparing Learning Processes Facilitated by ChatGPT and Human Experts
Early Career Research Grant

+ Enhancing Learning with Learning Analytics in an AI Powered World
+ March 2023 -> lots of challenges, potentials and unknowns about GAI
+ FLoRA project -> Self-regulation to Hybrid Human-AI Regulation
+ Exploratory, experimental, comparative research about ChatGPT

+ Note: we are comparing AI and human, but that’s not the main aim;
+ The aim is to better understand the relationships of AI and human.
Experimental design

• 133 participants (university students), AI in education, writing task
In this learning session, the goal is to write a vision essay that describes the future of education. Please describe, in 200 to 400 words, how you envision learning in a school in 2035.

Please consult the materials in the learning session for topics on envisioning the future.

Essay Writing

Normal 0 words

Write essay here...

For more information about the Essay Writing tool...

Checklist Tool

ChatGPT Tool

Ask Teacher Tool

(Stage 1)

Revise essay by themselves
Data collection

• Pre-Post test
• Pre-Post survey
• Learning trace data
• Eye-tracking and video
• Screen recording
• Dialogical text
• Post-study interview
• Etc.
Learning performance: essay scores

- Essay version 1 (after stage 1), essay version 2 (after stage 2)
- Essay scores improvement (essay version 2 ~ 1)

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Mean Difference</th>
<th>Lower Bound(95% CI)</th>
<th>Upper Bound(95% CI)</th>
<th>p-adjusted</th>
</tr>
</thead>
<tbody>
<tr>
<td>cl-ai</td>
<td>-2.200</td>
<td>-4.033</td>
<td>-0.367</td>
<td>0.012</td>
</tr>
<tr>
<td>cn-ai</td>
<td>-1.970</td>
<td>-3.858</td>
<td>-0.083</td>
<td>0.037</td>
</tr>
<tr>
<td>he-ai</td>
<td>-2.120</td>
<td>-4.049</td>
<td>-0.191</td>
<td>0.025</td>
</tr>
<tr>
<td>cn-cl</td>
<td>0.230</td>
<td>-1.725</td>
<td>2.184</td>
<td>0.990</td>
</tr>
<tr>
<td>he-cl</td>
<td>0.080</td>
<td>-1.915</td>
<td>2.075</td>
<td>1.000</td>
</tr>
<tr>
<td>he-cn</td>
<td>-0.150</td>
<td>-2.195</td>
<td>1.895</td>
<td>0.998</td>
</tr>
</tbody>
</table>

- AI group outperforms CN, CL and HE groups
Learning performance: knowledge gain and transfer

• Knowledge gain (pre-post-test): no significant differences
  • The ANOVA results show no significant differences between four groups in both the pre-test score (F=1.294, p=0.281, η²=0.036) and post-test score (F=0.913, p=0.438, η²=0.030).

• Transfer test (AI in medical science): no significant differences
  • ANOVA results show no significant differences between four groups (F=0.019, p=0.996, η²=0.000).
Intrinsic motivation (IMI, McAuley et al. 1989)

• No significant difference between the four groups was observed:
  • Interest/Enjoyment ($F=1.087$, $p=0.358$, $\eta^2=0.029$);
  • Perceived Competence ($F=0.453$, $p=0.716$, $\eta^2=0.012$);
  • Effort/Importance ($F=1.152$, $p=0.332$, $\eta^2=0.030$) and
  • Pressure/Tension ($F=0.546$, $p=0.652$, $\eta^2=0.015$).

• Although the insignificant were observed, we found:
  • CN group reported lowest interest and enjoyment, and highest pressure and tension
  • CL group reported highest scores for interest and enjoyment, perceived competence and effort, while they reported lowest pressure and tension
SRL processes

FLoRA trace parser

https://floraproject.org/website/

Identifying learning events in the learning environment

Defining SRL processes

Formulation of the SRL Model

<table>
<thead>
<tr>
<th>Main Categories</th>
<th>Subcategories</th>
<th>Codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metacognition</td>
<td>Orientation</td>
<td>MC.O</td>
</tr>
<tr>
<td></td>
<td>Planning</td>
<td>MC.P</td>
</tr>
<tr>
<td></td>
<td>Monitoring</td>
<td>MC.M</td>
</tr>
<tr>
<td></td>
<td>Evaluation</td>
<td>MC.E</td>
</tr>
<tr>
<td>Low_Cognition</td>
<td>First-reading</td>
<td>LC.F</td>
</tr>
<tr>
<td></td>
<td>Re-reading</td>
<td>LC.R</td>
</tr>
<tr>
<td>High_Cognition</td>
<td>Elaboration/Organisation</td>
<td>HC.E/O</td>
</tr>
</tbody>
</table>

The measurements of SRL with multi-channel data
Comparing learning processes of the first learning stage (2 hours reading and writing) among four groups.
Comparing learning processes of the second learning stage (1 hour revising) among four groups
Comparing process maps of revising stage between AI group and HE group

AI: stronger in red transitions
HE: stronger in green transitions

MC.O: Orientation
MC.P: Planning
MC.M: Monitoring
MC.E: Evaluation
LC.FR: First-Reading and Re-Reading
HE.EO: Elaboration and Organization

AI: Interact with ChatGPT
HE: Interact with Human Expert
What Experimental Study Tells Us? 1st Insight

+ High-intelligence tools (such as ChatGPT) may not stimulate intrinsic motivation to learn and knowledge gain/transfer, but can rapidly improve short-term performance;
+ Potential Metacognition Laziness and over-reliance, and the offloading of (meta)cognitive load can be the two sides of a coin;
+ “AI-empowered learning skills” which optimises performance at the expense of developing genuine human skills (?)
+ One important note: ChatGPT is excellent at utilizing clear rubrics

Fan et al., Beware of Metacognitive Laziness: Effects of Generative Artificial Intelligence on Learning Motivation, Processes, and Performance, British Journal of Educational Technology (under review)
Try to open this black box?

Other

Code the screen recording

Help-seeking process model
(Nelson-Le Gall, 1981)

<table>
<thead>
<tr>
<th>Stage</th>
<th>Activities</th>
<th>Definition</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnosing Question</td>
<td>Diagnosing Question</td>
<td>learners need to determine if there is a problem based on their own learning status and decide if they need help</td>
<td>Diag.Ques</td>
</tr>
<tr>
<td>Asking Help</td>
<td>Asking Help.Instrumental</td>
<td>Instrumental help-seeking (learners ask facilitated hints that can assist them in revising their essays independently afterward)</td>
<td>Ask.Instr</td>
</tr>
<tr>
<td>Asking Help</td>
<td>Asking Help.Executive</td>
<td>Executive help-seeking (learners tend to look for answers that can be applied directly)</td>
<td>Ask.Exec</td>
</tr>
<tr>
<td>Evaluating Help</td>
<td>Evaluating Help.Positive</td>
<td>Positive evaluation (learners evaluate the help messages and give positive feedback)</td>
<td>Eva.Pos</td>
</tr>
<tr>
<td>Evaluating Help</td>
<td>Evaluating Help.Negative</td>
<td>Negative evaluation (learners evaluate the help messages and give negative feedback or not to give any feedback)</td>
<td>Eva.Neg</td>
</tr>
<tr>
<td>Processing Help</td>
<td>Processing Help.Accepting</td>
<td>Accepting help (learners apply the help directly in revising their essays)</td>
<td>Pro.Acc</td>
</tr>
<tr>
<td>Processing Help</td>
<td>Processing Help.Neglecting</td>
<td>Neglecting help (learners do not apply the help directly in revising their essays)</td>
<td>Pro.Neg</td>
</tr>
<tr>
<td>Processing Help</td>
<td>Processing Help.Returning</td>
<td>Returning to the help (learners look back to some previous help message)</td>
<td>Pro.Re</td>
</tr>
</tbody>
</table>
Help-seeking

Try to open this black box?

Ask ChatGPT

Ask Teacher
## Comparison of activities between AI and HE Group

### Comparison of activities between AI Group and HE Group by Mann-Whitney $U$

<table>
<thead>
<tr>
<th>Activities</th>
<th>Mean in cognitive and behavioral activities (%) (AI, N=18)</th>
<th>Mean in cognitive and behavioral activities (%) (HE, N=20)</th>
<th>Mean Rank (AI, N=18)</th>
<th>Mean Rank (HE, N=20)</th>
<th>Z</th>
<th>Effect Size (ES)</th>
<th>Sig. (2-tailed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ask.Instr</td>
<td>64.75</td>
<td>77.86</td>
<td>22.83</td>
<td>16.50</td>
<td>-1.552</td>
<td>-0.183</td>
<td>.121</td>
</tr>
<tr>
<td>Ask.Exec</td>
<td>31.93</td>
<td>6.05</td>
<td>25.28</td>
<td>14.30</td>
<td>-3.141</td>
<td>0.756</td>
<td>.002**</td>
</tr>
<tr>
<td>Ask.Avo</td>
<td>3.30</td>
<td>16.08</td>
<td>17.00</td>
<td>21.75</td>
<td>-1.657</td>
<td>0.128</td>
<td>.098</td>
</tr>
<tr>
<td>Eva.Pos</td>
<td>2.22</td>
<td>54.55</td>
<td>10.83</td>
<td>27.30</td>
<td>-4.937</td>
<td>-0.778</td>
<td>.000***</td>
</tr>
<tr>
<td>Eva.Neg</td>
<td>97.77</td>
<td>45.44</td>
<td>24.22</td>
<td>15.25</td>
<td>-4.937</td>
<td>0.778</td>
<td>.000***</td>
</tr>
<tr>
<td>Pro.Acc</td>
<td>64.92</td>
<td>60.41</td>
<td>19.58</td>
<td>19.43</td>
<td>-1.142</td>
<td>0.256</td>
<td>.253</td>
</tr>
<tr>
<td>Pro.Neg</td>
<td>8.42</td>
<td>3.24</td>
<td>23.53</td>
<td>15.88</td>
<td>-2.231</td>
<td>0.678</td>
<td>.026*</td>
</tr>
<tr>
<td>Pro.Re</td>
<td>26.65</td>
<td>36.33</td>
<td>15.36</td>
<td>23.23</td>
<td>-2.373</td>
<td>-0.417</td>
<td>.018*</td>
</tr>
</tbody>
</table>
General Instructions

In this learning session, the goal is to write a vision essay that describes the future of education. Please describe, in 200 to 400 words, how you envision learning in a school in 2035.

Please consult the materials in this learning environment that provide information about three important topics for envisioning the future of education in 2035.

1. Artificial intelligence and its application
2. Differentiation and how it is applied in the classroom
3. The process of scaffolding and how it optimizes students learning

The goal of the learning session is to integrate these topics into a school in 2035.

At the end of the learning session, you should be able to:
- explain the concepts of artificial intelligence, scaffolding and differentiation strategies. This could result in a more inclusive and effective learning environment, where all students are able to reach their full potential.

For more information about the criteria of the essay, you are required to read the rubric.

In this session, you will revise the essay in 60 minutes. Please make sure to work efficiently. We advise you to focus on the rubric.
Dialogical interaction and mechanism

"Could you please evaluate/provide feedback on my essay?"

"give me feedback on my essay, or Please better link these two sentences"

"how can I make my essay formal?" (Seek information about steps or procedures in a process)

Epistemic Network Analysis
AI: stronger in red edges
HE: stronger in blue edges
What Experimental Study Tells Us? 2nd Insight

+ Learners ask AI pragmatic questions -> improve their performance;
+ Learners experienced lower social cost in the AI group compared to asking the human expert;
+ Learners showed adaptivity when facing different facilitators;
+ Previous theories and models (e.g., Linear help-seeking process) may encounter difficulties in explaining human-AI interaction;
+ Concerns about learners’ lack of evaluation and monitoring when seeking executive help from ChatGPT -> scaffoldings

Chen et al., Unpacking Help-Seeking Processes through Multimodal Learning Analytics: A Comparative Study of Learning Facilitated by ChatGPT and Human Expert (drafting);
Cheng et al., Asking Questions of Generative Artificial Intelligence Improves Academic Performance (drafting)
But, who did learners prefer to learn with?

- Human-AI preference scale (5 questions, pre and post task)
  - e.g, “Compared to AI, human tutors can better understand the main idea of my article and provide more helpful suggestions.”
- Human-AI choice (pre and post task)
  - 1 -> prefer human over AI; 0 -> prefer AI over human
- Preference Alteration
  - -1 indicates participant altered preference from human tutor to AI;
  - 1 indicates participant has altered preference from AI to human tutor;
  - 0 indicates participant’s preference remains unchanged.

<table>
<thead>
<tr>
<th>Pre-test 30 mins</th>
<th>Training 10 mins</th>
<th>STAGE1: Reading and Writing 120 mins</th>
<th>Training 10 mins</th>
<th>STAGE2: Revising 60 mins</th>
<th>Post-test 30 mins</th>
</tr>
</thead>
</table>

1st preference measure

2nd preference measure
Human-AI preference

No difference before task
In general, learners prefer human

Preference Alteration Score by Group

changed preference from AI to human tutor
changed preference from human tutor to AI
Polarization of preference change

- AI group who has experienced ChatGPT interaction are more inclined to AI
- HE group who has experienced human expert interaction are more inclined to humans (100% choose humans)
- CN group shows no significant change, very slightly turns to human
- CL group who has not experienced Chatgpt and human expert, surprisingly, also are more inclined to humans (97% choose human)
<table>
<thead>
<tr>
<th>Principle</th>
<th>Format of feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic</td>
<td>Highlight spelling grammar errors</td>
</tr>
<tr>
<td>Check based on GPT4.0</td>
<td></td>
</tr>
<tr>
<td>Academic</td>
<td>Highlight Words/phrases with problematic academic style</td>
</tr>
<tr>
<td>Based on a database of academic norms developed by teachers</td>
<td></td>
</tr>
<tr>
<td>Originality</td>
<td>Highlight repeat sentences</td>
</tr>
<tr>
<td>Calculation based on similarity of more than 7 words</td>
<td></td>
</tr>
<tr>
<td>Integration and Elaboration</td>
<td>Different color highlights represent different levels of processing</td>
</tr>
<tr>
<td>Division based on Bloom’s taxonomy of cognitive domains</td>
<td></td>
</tr>
</tbody>
</table>
What Experimental Study Tells Us? 3rd Insight

+ ChatGPT and Checklist: what is the main difference?
+ The dialogic format may be naturally more attractive than traditional feedback tools or dashboard;
+ The intelligence level of AI agents or learning tools affects learners’ trust and aversion on algorithms/techniques.
+ No matter how the data show the practical of learner-AI interaction, human (teachers) always have irreplaceable human’s value, and learners STILL prefer to learn with human expert in our task.

Le et al., Rolling to the edge: investigating learners’ preference for learning supports from human-tutor, traditional AI tutor and LLM (drafting)
Interviews to understand learners’ perspective

AI group: 33 accepted the interview
HE group: 26 accepted the interview
Different values and value alignment

<table>
<thead>
<tr>
<th>Value</th>
<th>Working definitions</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Empathy</td>
<td>comprehend and respond to learner’s intentions and emotions properly</td>
<td>AI27: Even if my question was not clear enough, the teacher still knew what I wanted to ask.</td>
</tr>
<tr>
<td>Care</td>
<td>alleviate the interpersonal stress of learners and demonstrate patience to them</td>
<td>HE03: I hope the teacher will never get tired and judgmental about my questions.</td>
</tr>
<tr>
<td>Autonomy</td>
<td>Respect for learners’ freedom to make their own choices and to monitor their learning process</td>
<td>AI13: I have greater power of selectivity and decision-making and greater freedom.</td>
</tr>
</tbody>
</table>

Interview Data

Better embodied value by teacher (human expert)
Better embodied value by AI (ChatGPT)
Value tensions of learning with AI or human expert

Inter-stakeholder Value Tensions

**Developers' Value**

**Researchers' Value**

**Teachers' Value**

**Learners' Value**

**Ethical Value**
- Empathy
- Autonomy
- Freedom from bias
- ...

**External Tension**
- embodying one ethical value may inhibit another pragmatic value, vice versa

**Internal Tension**
- embodying one ethical value may inhibit another ethical value, vice versa

**Conceptual Tension**
- different interpretations of a value’s meaning lead to divergent priorities

**Pragmatic Value**
- Development
- Accuracy
- Efficiency
- ...

...
What Experimental Study Tells Us? 4th Insight

+ Human and AI (e.g., ChatGPT) each have their own unique value;
+ Learners also dynamically perceive and evaluate affordances of different learning facilitators as they regulate their own learning;
+ Value as a key ethic issue of AI in education was relatively neglected;
+ Different stakeholders should keep value sensitive design in mind and seek the balance between different values;

Shen et al., Aligning and Comparing Values of ChatGPT and Human as Learning Facilitators: a Value-Sensitive Design Approach, British Journal of Educational Technology (under review)
Let’s get back to the CL group.
Trigger and adjust self-assessment

**Experimental procedure**

**STAGE1: Reading & Essay Writing (2 hours)**
- Group CL & CN
  - Reading Material
  - Rubric
  - First Essay
  - Self-Assessment Score 1
  - Mini Interview 1
  - Researcher Assessment Score 1

**STAGE2: Essay Revision (1 hour)**
- Group CL
  - Reading Material
  - Rubric
  - Checklist tool
  - Final Essay
  - Self-Assessment Score 2

- Group CN
  - Reading Material
  - Rubric
  - Mini Interview 2

**Data processing**

- Self-Assessment Score 1 & 2
- Researcher Assessment Score 1 & 2

- Spearman correlation: Self-Assessment Score & Researcher Assessment Score
- T-test: | Self-Assessment Score - Researcher Assessment Score |
- Distribution: Self-Assessment Score - Researcher Assessment Score

- The accuracy of self-assessment on learning products

**Coding**
- Overall performance
- Satisfaction & Dissatisfaction
- Influence factor
- Plan
- Experience

- The process of self-assessment
Checklist tools improved self-assessment, and performed better than CN (and AI and HE) groups
Factors that influenced writing: learners’ feedback

CL group

CN group
What Experimental Study Tells Us? 5th Insight

+ Writing analytics feedback tools or dashboards also have unique values;
+ Adjusting learners’ self-assessment is a core part of self-regulated learning and has been under-researched in previous studies;
+ Low affordability tools (such as bloom taxonomy tool in the Checklist toolkit) are not very useful for self-regulated learning;
+ When tools provided, regardless of their affordance, learners will subjectively consider tools as the primary factor, thereby inhibiting reflection on their own abilities.

Tang et al., Facilitating Learners' Self-assessment during Formative Writing Tasks using Writing Analytics Toolkit, Journal of Computer Assisted Learning, 2024 (accepted)
Discussion: Human and AI (人和人工智能)

At the 2008 Beijing Olympics Opening Ceremony, thousands of actors spent five minutes performing only one Chinese character: 和 (hé)

Conjunction or preposition: and, with

Noun: sum, peace, etc

Verb: mix, agree, join, blend, fellow, etc

Adjective: gentle, moderate, harmonious
Discussion: the rich connotations of “和”

- Learning and regulating with ChatGPT, but NOT simply using AI as a tutor to replace human teacher;
- Learning followed, joined and moderated by AI;
- Future learning and teaching of combining human and AI;
- Hybrid intelligence: learner-AI, teacher-AI, learner-teacher-AI, etc;
- The interplay of human control and AI automation (Cukurova, 2024);
- Scaffolding high human and AI-empowered skills (Gasevic, 2024);
- Different stakeholders learn, work and live in harmony with artificial intelligence!
Future works: AI-scaffolded dialogue space
I'm hiring!

Please consider join our team if you are looking for post-doc position and find our studies interesting! email -> fyz@pku.edu.cn
Thank you!

Yizhou Fan
fyz@pku.edu.cn
Peking University