Do We Learn From Each Other: Understanding the Human-AI Co-Learning Process Embedded in Human-AI Collaboration

Do We Learn From Each Other: Understanding the Human-AI Co-Learning Process Embedded in Human-AI Collaboration

Do We Learn From Each Other: Understanding the Human-AI Co-learning Process Embedded in Human-AI Collaboration

Jinwei Lu[1], Yikuan Yan[1], Keman Huang[1,2*], Ming Yin[3], Fang Zhang[4]

[1]Renmin University of China, No. 59 Zhongguancun Street, Haidian District, 100872, China.
[2]MIT Sloan, 100 Main St, Cambridge, MA 02143, USA.
[3]Purdue University, 305 N. University St., West Lafayette, IN, USA, 47907.
[4]Tsinghua University, 30 Shuangqing Rd, Beijing, 100086, China.

*Corresponding author(s). E-mail(s): keman@ruc.edu.cn;
Contributing authors: lujinwei2025@qq.com; yanyk_cufe2019@163.com; mingyin@purdue.edu; zhangfang2021@tsinghua.edu.cn;

1

Do We Learn From Each Other: Understanding the Human-AI Co-learning Process Embedded in Human-AI Collaboration

Abstract

Beyond collaborating in the AI-supported decision-making setting to achieve complementary performance, human and AI should learn from each other and internalize knowledge from their collaboration. This can enhance their individual performance when working independently after their collaboration. However, this expected dual-pathway co-learning process, including both *"human learns from AI"* and *"AI learns from human"*, does not occur spontaneously. Human-AI collaboration designs could have inconsistent and intertwined influences on the co-learning process. Based on the learning cycle theory, this study conducted three online, two-stage, and between-subject behavioral experiments to reveal how human and AI learn from each other. By developing a context where human and AI have comparable and moderate performance on emotion classification tasks, our study provides the first empirical evidence of an effective human-AI co-learning process within human-AI collaboration. However, the AI feedback and collaborative workflow design can lead to unequal and potentially negative impacts on both pathways of the co-learning process in groups with varying levels of cognitive reflection capability. These findings highlight three design principles to facilitate the co-learning process embedded in human-AI collaboration rather than naively deploying a complex AI system.

1 Introduction

Recent studies have investigated how human and AI collaborate together to achieve complementary performance that exceeds human or AI alone [1–4]. However, the result is mixed at best [3, 5], motivating many studies to investigate factors that can affect the effectiveness of the human-AI collaboration, such as communication method [6], task delegation [7], decision-making processes [4] and the AI's identity [6, 8, 9]. It has become a consensus that to achieve complementary performance between human and AI, AI should be regarded as a partner [10] rather than merely a support tool. It is essential to harness and leverage the complementary knowledge of both human and AI partners [2–4, 11]. In other words, human and AI should be able to learn from each other through their collaboration, making it particularly important to explore such a human-AI co-learning process.

However, while emphasized in human-human collective intelligent studies [12, 13], our understanding of this co-learning process embedded in the human-AI collaboration is very limited [14, 15]. As shown in Figure 1, the co-learning process is a dual-pathway that both human and AI adapt to each other and learn through their collaborative experience [14]. It includes human digesting the results from AI for learning — *"human learns from AI"*, and AI being trained based on human's inputs — *"AI learns from human"*[1]. Ideally, human-AI collaboration can complement each other through this co-learning process, leading to a closer mental model and achieving not only complementary performance when collaborating together but also better performance when working alone after their collaboration [15–17] (a.k.a, the *training phase* and *testing phase* elaborated as the left and right part of Figure 1 separately). However, most existing studies have focused on the human-AI collaboration in an AI-supported decision-making setting, aiming at achieving complementary performance when working together [10, 18]. It is essential to further explore how such a human-AI collaboration experience affects their individual performance after their collaboration to guide designs that can enhance their co-learning effectiveness [19, 20].

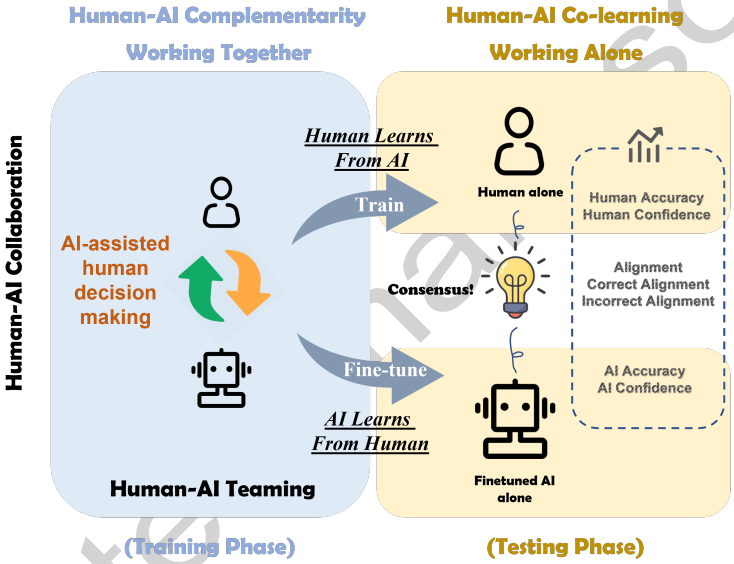


Fig. 1: The dual-pathway human-AI co-learning process within their collaboration.

Importantly, the Kolb's learning cycle theory [14, 15, 20–23], have served as the foundation for understanding the learning process and assessing how distinct learning processes influence the acquisition of various types of knowledge. It highlights a learning process involving multiple stages including concrete experience (being involved in

[1]Notably, rather than taking an active learning paradigm which makes two pathways of the co-learning process intertwined, we adopt a training-testing two-stage design that we use the generated annotations from the training phase to finetune the AI to represent the AI learning from human process.

a new experience), reflective observation (reflecting on the observation from the experience), abstract conceptualization (creating theories from the observation) and active experimentation (using theories to solve new problems) [21]. Learners go through this process to create and internalize knowledge from their experience and then apply what they learn in real situations [20]. Notably, the transformation from one stage to another is not naturally occurring but can be influenced by many factors [14, 21, 22]. Hence, in the human-AI collaboration context, the co-learning process, including both pathways, also follow such a process that the learning may not happen spontaneously [17]. Although few recent studies on human-AI collaboration have started to notice this aspect [17, 24–27], reporting some scattered observations, there is still a lack of theoretical frameworks and clear empirical evidences to unfold this co-learning process.

Therefore, the goal of this study is to investigate the human-AI co-learning process embedded in human-AI collaboration based on the Kolb's learning cycle theory. In particular, as suggested by the Kolb's learning cycle theory [21, 23], concrete experience is the central and necessary part of the learning process, which triggers the learning cycle. Moreover, learners' collaboration is essential for success of this process [28]. Given that the human-AI collaboration creates a collaborative learning experience for both human and AI, the first question this study explores is:

- **RQ1**: *Can the human-AI collaboration experience trigger the dual-pathway co-learning process that improves their individual performance when working alone after the collaboration?*

Additionally, the reflective observation stage, whereby the learner reflects on the observations, acting as a mediator of meaning making, is imperative in the learning process [21, 23]. Facilitating this stage to cognitively engage learners into deeper thinking, a.k.a "learning by thinking", will benefit in acquiring knowledge and improve the learning outcome [22]. In the human-AI collaboration context, AI feedback designs [29] and collaborative workflow configurations [11, 17, 30] have been proposed to enhance human's cognitive engagement with AI to promote their collaboration. In particular, AI feedback designs provide richness of information provision, especially the rationale behind AI's decision. The collaborative workflow configurations enhance the extent of human cognitive participation through double and slow thinking. They offer different mechanisms for enhancing cognitive engagement between human and AI. Hence, this study aims to investigate how these two different cognitive engagement mechanisms affect the co-learning process:

- **RQ2**: *How can different AI feedback designs in human-AI collaboration influence the dual-pathway co-learning process?*
- **RQ3**: *How can different collaborative workflow designs in human-AI collaboration influence the dual-pathway co-learning process?*

Finally, in the abstract conceptualization stage, learners can create concepts that integrate their observations into logically sound theories, which transforms what they observed into knowledge [20, 21]. Existing studies following the dual process theory highlight that cognitive reflection capability, the way people engage with provided information when making decisions [31, 32], is a critical internal factor for learning.

People with a higher cognitive reflection capability will achieve better learning performance when abstracting concepts from their experience, suggesting the importance of adaptive learning [33, 34]. Particularly, existing studies investigating AI feedback and collaborative workflow designs demonstrate that people with different cognitive reflection capabilities can engage with the AI-generated information differently, indicating the intervention-generated inequality effect [17, 35, 36]. For example, people with a high cognitive reflection capability can engage deeper and benefit more from the correct and well-designed AI explanations [17]. Hence, when designing human-AI collaboration, it should account for the human's cognitive reflection capabilities [4, 32, 37]. This motivates us to further explore:

- **RQ4**: *How does the cognitive reflection capability in human-AI collaboration affect the dual-pathway co-learning process?*

To answer these questions, we conduct three online, two-stage, and between-subject behavioral experiments. More specifically, to answer **RQ1**, in Study 1, participants were asked to complete 32 multi-category emotional annotation tasks[2] with or without AI recommendations in the training phase. In the testing phase, participants will complete another 32 tasks without AI support while the annotations from the training phase are used to fine-tune the AI models, and the fine-tuned AIs are used to predict the tasks. We further propose measures including the *human accuracy and confidence*, *AI accuracy and confidence* as well as their *alignment* and *(in)correct alignment* in the testing phase (as summarized in the right part of Figure 1) to evaluate both pathways of the co-learning process. This enables us to quantitatively investigate whether the human-AI collaboration experience in the training phase triggers an effective co-learning process to improve their performance when working alone in the testing phase. After confirming the existence of the human-AI co-learning process from Study 1, following the previous studies (i.e. [17, 38, 39]), we answer **RQ2** by conducting Study 2 to investigate how different feedback designs can impact this process. In particular, we consider three feedback conditions, including *providing AI recommendation only*, *providing AI explanation only*, or *providing AI recommendation and explanation simultaneously*, which provides more AI feedback rationales sequentially to engage human during the training phase. This will allow us to understand how providing additional AI feedback affects the co-learning process. To answer **RQ3**, in Study 3, we further design a two-step collaborative workflow for human-AI interaction, where participants can update their annotations after viewing the AI recommendation [11, 17], to understand how the collaborative workflow configuration impact the co-learning process. Finally, we conduct an internal audit to split our data by human cognitive reflection capability. This aims to investigate the potential intervention-generated inequality

[2]Emotional annotation of posts is a human instinct that requires no specific training to get started. It is also a nontrivial task that human cannot judge straightforwardly, allowing us to build a learning environment where human and AI have both imperfect but complementary knowledge of the given tasks. Existing studies showcasing complementary performance in human-AI collaboration emphasize making the complementary region between human and AI's knowledge space larger to achieve optimal outcome [4, 11]. Hence, in this study, we design a context where human and AI have moderate and comparable (both around 50% accuracy) performance so that neither human nor AI exhibit a significant advantage or disadvantage, leaving room for them to learn from each other if the co-learning process can happen.

effect in the human-AI co-learning process due to the cognitive reflection capability difference, which answers **RQ4**.

To the best of our knowledge, this is the first study to provide clear empirical evidence of the dual-pathway human-AI co-learning process embedded in human-AI collaboration and how the human-AI collaboration designs impact this co-learning process. In our study, where human and AI have comparable and moderate accuracy, we find that the human-AI collaboration experience can trigger both pathways of the co-learning process, confirming the existence of human-AI co-learning [17, 24–27] and highlights the importance of collaborative experience for learning [21, 23]. However, AI feedback fails to consistently enhance both pathways, indicating an inequality effect between the two learning processes. Similarly, the two-step collaborative workflow design fails to stimulate better co-learning performance for both participants. Rather, it can result in some negative and unequal impacts on both pathways. These impacts also vary based on human cognitive reflection capabilities. Therefore, naively deploying a complex AI system without optimizing the design of AI feedback and collaborative workflow could not only fail to enhance their collaboration [4, 31, 40–43], but also negatively affect the co-learning process. Moreover, going beyond existing studies [17, 35, 36], our study shows that the intervention-generated inequality effect exists in both groups with different human cognitive reflection capabilities and within the two co-learning pathways. Hence, when designing to facilitate both pathways in the human-AI co-learning process, the top priority is to create a collaborative team between AI and human, emphasizing that *"something is better than nothing"*. However, the design should balance the generated information and the collaboration workflow, which may harm the co-learning outcome. In other words, *"more can be less"*. Moreover, participants with different cognitive reflection capabilities will impact both pathways in the co-learning process unequally, underscoring the importance of adaptive learning oriented design, a.k.a, *"one size can not fit all"*. Overall, by extending existing studies to investigate the co-learning process, we provide valuable guidance to promote human-AI learning, especially as AI is increasingly considered a self-learning, proactive, and intelligent subject [31, 44, 45].

In the following, we first review related work and then propose our hypothesis. After detailing our three studies, including the experimental design, participant demographic, and observations from the behavioral experiments, we conduct the internal audit to uncover the intervention-generated inequality phenomenons. Finally, we summarize our findings and discuss the limitations and future directions, followed by a final remark to conclude this study.

2 Related Work

2.1 The Co-Learning Process within Human-AI Collaboration

A growing body of research has sought to design effective human-AI collaboration that human and AI systems can compensate for each other to achieve complementary performance [2–4, 6, 8, 46]. Typically, AIs are considered as a decision-support tool [3, 10], wherein it collaborates with human to accomplish specific tasks. However, although some studies demonstrate a complementary performance [2, 3, 47], many studies also

report a failure to improve the collaborative performance [30]. For the human-AI collaboration which outperforms people working alone [4, 30], the improvement can be due to the setting that AI's accuracy is significantly higher than human initially [4, 48] as the collaborative outcome rarely outperforms AI alone [5]. Hence, exploring the collaborative mechanisms within human-AI collaboration is still an open but critical research area [49].

Importantly, inspired by human-human collective intelligence studies [12, 13], some recent studies view human and AI as equal entities, considering AI as a partner within a human-AI team [6, 8, 18, 50] rather than mere a tool. Illustrative examples of such teamwork can be found in various domains, such as healthcare applications [51], recidivism risk assessment [52], computer games [53], and educational settings [54]. In particular, when conceptualized as a teammate, AI takes on team-level responsibilities and executes decisions as a team member to contribute to shared team goals [53]. AI is considered as a self-learning, proactive, and intelligent subject [44, 45]. In other words, AI is a partner rather than just a support tool, so building shared mental model [10] between human and AI becomes critical for their effective collaboration. Hence, the co-learning process [15, 55, 56], a process where partners adapt to and learn from each other to establish the shared mental models on the task, themselves and their partners, plays an increasingly crucial role. For example, Zagalsky et al. [57] proposed a human-machine learning (HML) framework with reciprocal learning feedback to enable "learning-by-doing" where domain experts and AIs learn simultaneously and incrementally to reach a shared conceptualization.

Notably, the co-learning process for human-AI collaboration is a dual-pathway learning process, which involves *"human learns from AI"* by digesting the results of AI and *"AI learns from human"* through training based on human's input [15]. Ideally, it will form a collaborative learning environment that improves both participants' performance, resulting in better collaborative performance and performance improvement for both human and AI when working alone after their collaboration. Unfortunately, such a process does not occur spontaneously [17, 58]. For example, Schoonderwoerd et al. [58] use the restricted-perception Wizard-of-Oz (WOz) approach to simulate a human-AI collaboration for an urban search and rescue mission. However, no learning effects on collaborative fluency or performance were found. Similarly, using a nutritional knowledge quiz task, Gajos and Mamykina [17] show that providing AI recommendations and explanations can not achieve learning benefits as it does not involve human in the cognitive processing of AI-generated information.

As suggested by the learning cycle theory [20–23], learning is a dynamic knowledge internalization process that involves multiple steps, including observation, comprehension, application, and the accumulation of experience [20]. Learning is a cyclical process wherein individuals continuously reflect on and reapply knowledge through practical activities [20, 23]. This cycle not only aids individuals in understanding new information but also enables them to internalize this information into personal tacit knowledge and apply it in new situations [20, 23]. However, this transformation will not happen spontaneously but is affected by many factors, like collaborative experience, communication mechanism and critical reflection [14, 21, 22]. Hence, in the

context of human-AI co-learning, it is important to consider not only the tasks completed jointly but also the performance of tasks completed independently by human and AI to evaluate the effectiveness of their collaboration, which is often overlooked.

Overall, while existing studies recognized the essentials to enable the learning process to achieve complementary performance, the two learning pathways do not happen spontaneously and they are not effectively distinguished. Given the lack of empirical evidence, this study aims to investigate how human-AI collaboration will impact the co-learning process.

2.2 The Intertwined Factors Affecting Human-AI Collaboration

As discussed above, many recent studies [4, 10, 18, 32] have reported that many factors can influence human-AI collaboration. Among these factors, the human-AI collaboration designs in AI feedback and collaborative workflow have been considered critical, as they provide different mechanisms to engage human in their collaborations.

First, the AI feedback design focuses on providing explanations regarding AI's choice to support human's decision during the human-AI collaboration process. It aims to help users understand how AI systems work [38, 59] to improve user's trust and satisfaction [38], and encourage the use of AI systems [26]. Many studies have confirmed the effectiveness of explanations for AI decisions to support human-AI collaboration, including professional scenarios like healthcare [51]. However, the complexity of explanation [60], user expectations [37, 61], faithfulness [41] can make the explanation much less effective and sometimes can even result in adverse outcomes due to overreliance effect, anchor bias, etc. [3, 4]. For example, Lim and Dey [61] found that extensive explanations increase the perceived accuracy of high-precision systems but decrease the perceived accuracy of low-precision systems. Papenmeier et al. [42] demonstrated that providing faithful or random explanations in high-precision systems reduces users' self-reported trust. Smith-Renner et al. [43] showed that adding explanations to low-precision models decreased users' acceptance and confidence in the model and increased frustration.

Second, designing the collaborative workflow between human and AI, which provide cognitive forcing strategies such as double-thinking or slow decision making, is proposed to reduce the overreliance and improve the human-AI collaboration [17, 30, 48]. For example, the two-step workflow design that requires users to decide before seeing the AI's suggestion [48] allows users to choose whether and when to see the AI's suggestion, providing a double thinking strategy to engage human participants during the collaboration. However, these cognitive forcing strategies at the cost of the extra time [31] can increase the mental effort required for decision-making, which can cause unintended consequences [62].

Moreover, both designs aim at enhancing the cognitive engagement with human during the human-AI collaboration. When humans engage in extended thinking and learning processes, they often tend to think and learn analytically, which is a manifestation of cognitive reflection capability [63]. In particular, the dual process theory in psychology suggests that biases in human decision making may be explained by the type of cognitive processes that people engage in when making decisions—"System 1"

processing is executed quickly without reflection, while "System 2" requires conscious thought and effort [31]. An individual tending to engage with conscious processing ("System 2") more often, indicated as a high cognitive reflection capability [31], naturally exhibits a more analytical disposition to digest the provided information. Hence, when designing the human-AI collaboration mechanisms, this could result in intervention-generated inequalities [17, 32, 37]. For example, Gilpin et al. [37] suggest accounting for the limited human attention span and cognitive reflection capabilities when designing explanations, e.g., by showing what brought about a single decision rather than explaining the machine learning model as a whole. Duan et al. [32] demonstrate that people's high cognitive reflection capability could mediate the influence of task designs on recidivism risk judgments. Gajos et al. [17] highlight the AI explanation benefits people with high cognitive reflection capability more.

In summary, existing studies have extensively investigated designs such as AI feedback and collaborative workflow to cognitively engage human to improve human-AI collaboration, as well as the intervention inequality arising from variations in cognitive reflection capability. However, the inconsistent observations call for more empirical evidences regarding their performance to support human-AI collaboration. More importantly, these factors provide different mechanisms to engage humans with different cognitive capabilities in their collaboration, which could affect the learning process as suggested by the learning cycle theory [14, 21, 22]. Hence, going beyond existing studies that focus on the AI-supported decision making setting, we investigate how these designs intertwined together to affect the co-learning process.

3 Hypothesis Development

As shown in Figure 2, to investigate the human-AI co-learning process, the first goal is to verify *whether the co-learning process occurs within the human-AI collaboration* to answer **RQ1**. In particular, as suggested by the learning cycle theory [14, 20–22], bringing a community together plays a key role in an effective learning experience because learning is a social and collaborative process [64]. The learner's communication is essential for the success of the process [28] because they can learn with and from each other [65]. Existing studies have demonstrated that involving AI to form a collaborative team could achieve a better performance compared with human alone, indicating an effective collaboration experience during the training phase [1–4, 48]. Hence, we could expect that involving AI to form a collaborative team with human should be able to trigger a learning process, which improves human performance when working alone in the testing phase.

Additionally, in the human-AI collaboration context, the learning process is a dual-pathway learning process, which involves human learning from AI by digesting the AI' feedback and AI learning from human through training using their collaborative outcomes [15, 57]. If an effective co-learning process occurs, the human-AI collaboration would foster effective communication between human and AI, enabling the human to understand and apply the AI's suggestion to generate valuable collaborative outcomes that the AI can learn from [15]. Hence, after fine-tuning using this outcome, AI could also perform better when working alone in the testing phase.

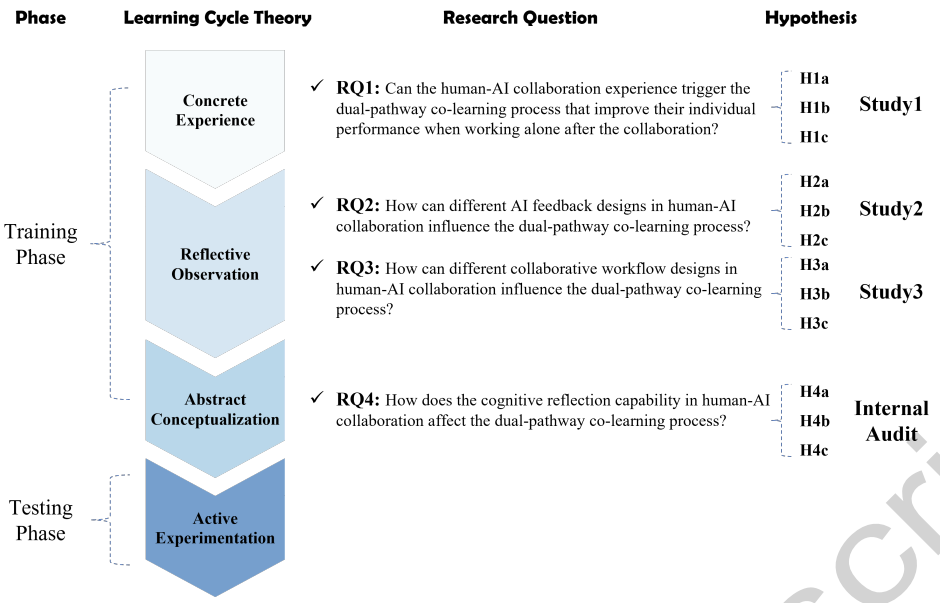


Fig. 2: Research questions based on learning cycle theory.

Moreover, as both human and AI learn from each other through the training phase, human can understand and apply the analyses of the AI and AI can adapt from human cognition, thereby achieving a more aligned mental model [15]. Hence, given the same tasks in the testing phase, they should perform more similarly even when working alone. Therefore, we can propose the following hypotheses:

- **H1a**: Human can learn from the human-AI collaboration experience, leading to an improvement in the individual performance when operating independently after the collaboration.
- **H1b**: AI can learn from the human-AI collaboration experience, leading to an improvement in the individual performance when operating independently after the collaboration.
- **H1c**: After experiencing the human-AI collaboration, human and AI become more aligned with each other, achieving a closer mental model.

Critical reflection is imperative in the learning process [23]. When people are cognitively engaged with instructional processes and materials, they are more likely to benefit from instruction and are more likely to acquire new skills and knowledge. In particular, compared with learners simply observing the provided materials (a.k.a, *know-what*), understanding the cause-effect relationships behind those materials (a.k.a, *know-why*) can encourage more critical reflection to heighten the learning effectiveness [20, 22]. Hence, in the context of human-AI collaboration, through providing AI's recommendations accompanying explanations, human could get an insight into how AI arrived at its recommendations, therefore achieving a better learning outcome [17]. Additionally, the explanations were supposed to help human identify the

situations in which AI's reasoning was incorrect, resulting in more correct collaborative output, especially in complex situations where annotators lack contextual information or complete domain knowledge [66]. This can then be used to fine-tune AI to correct its model to improve its performance, as well as become more aligned with the human mental model in the testing phase. Therefore, we can propose the following hypotheses to answer **Q2**.

- **H2a**: Providing additional explainable information through AI feedback designs can enhance the human learning process, leading to an improvement in the individual performance when operating independently after the collaboration.
- **H2b**: Providing additional explainable information through AI feedback designs can enhance the AI's learning process, leading to an improvement in individual performance when operating independently after the collaboration.
- **H2c**: Providing additional explainable information through AI feedback designs can further lead to closer human-AI mental models.

Existing studies demonstrated that people might over-rely on the AI-generated recommendations [4] as they did not really engage with the explanations but simply proceeded with the action. Simply presenting a person with a decision suggestion and an explanation would provide an immediate benefit (i.e., help the person make a better decision) but may not lead to learning [17]. Recent work demonstrated that certain kinds of in-the-moment interventions—such as forcing people to wait before submitting a decision, having people state their initial decision before seeing the AI recommendation, or having the AI recommendation provided only on demand—can reduce (but not eliminate) this overreliance and enhance the cognitive engagement [48]. Specially, a two-step collaborative workflow, which allows participants to update their initial decisions after being presented with the AI recommendation [48], can improve human-AI collaboration. One reason that these interventions are effective is that they provide a double-thinking stage to encourage people to engage more deeply with the AI-provided information, representing an effective reflection and analysis, which is the critical stage for the learning process [67, 68]. Additionally, such a deeper cognitive engagement, where they critically examine these recommendations and compare them with their own knowledge and judgment, would enable learners to reject information that is inconsistent with individuals' own knowledge and beliefs and to adjust their own knowledge to incorporate new information. Therefore, we can propose the following hypotheses to answer **Q3**:

- **H3a**: Providing double thinking strategy through collaborative workflow designs can enhance the human learning process, leading to an improvement in individual performance when operating independently after the collaboration.
- **H3b**: Providing double thinking strategy through collaborative workflow designs can enhance the AI's learning process, leading to an improvement in the individual performance when operating independently after the collaboration.
- **H3c**: Providing double thinking strategy through collaborative workflow can further lead to closer human-AI mental models.

Notably, the learning process is a cognitively engaging process. Existing research indicates that "learning by thinking" benefits in acquiring knowledge and improves learning outcomes [22]. Therefore, the cognitive reflection capability of human is particularly important. Individuals with higher cognitive reflection capabilities, who predominantly employ "System 2" in their cognitive processes [31, 69], are better able to pause and think when faced with problems requiring the overcoming of intuitive incorrect answers, thus performing better on tasks involving critical thinking and reflection [31]. Consequently, in the context of human-AI collaboration, humans with higher cognitive reflection capabilities tend to critically evaluate AI feedback [70], leading to higher performance and less prone to misinformation. These high-quality answers serve as human input to correct the AI model and improve its performance. Hence, similarly, this could result in more aligned mental models between human and AI. Therefore, we can propose the following hypotheses to answer **Q4**:

- **H4a**: Human with higher cognitive reflection capability in collaboration team can enhance the human's learning process, leading to better individual performance for human when operating independently after the collaboration.
- **H4b**: Human with higher cognitive reflection capability in collaboration team can enhance the AI's learning process, leading to better individual performance for AI when operating independently after the collaboration.
- **H4c**: Human with higher cognitive reflection capability in collaboration team could lead to a closer human-AI mental models.

To verify these hypotheses, in the following sections, we detail our three online behavioral experiments and then report our internal audit regarding the potential intervention-generated inequality because of differences in human cognitive reflection capabilities.

4 Study 1: The existence of co-learning process

The first study aimed to explore *whether the co-learning process occurs within the human-AI collaboration* to answer **RQ1**. According to above proposed hypothesis, if an effective co-learning process happens, the human-AI collaboration experience during the training phase should improve human and AI performance when they work alone in the testing phase.

4.1 Task Selection

In this study, we choose the multi-category emotional annotation task, where participants classified a given social media post into one of seven emotional types: *anger, disgust, fear, happiness, like, sadness, and surprise*. We use the dataset from the NLPCC 2014: Emotion Analysis in Chinese Weibo Texts, which is provided by Sina Weibo in 2014. We choose this task because (1) the multi-category emotional annotation task is generally regarded as a natural language understanding task that humans can accomplish without specific training or domain knowledge, and advanced AI models have been developed to simulate human capabilities. Therefore, both human and AI can have the ability and confidence to undertake it. Additionally, (2) the task

requires classifying each post into one of seven emotion types. Hence, it is non-trivial for humans to judge straightforwardly but engage necessary cognitive efforts.

4.2 AI Model and Sample Selection

We choose the ChineseMacBert model [71], a well-known pre-trained deep learning model for Chinese natural language processing (NLP), which has achieved state-of-the-art performances on many Chinese NLP tasks. Then, we use the 6149 samples with emotion labels from the training dataset to fine-tune a ChineseMacBert model. For the convenience of expression, in this study, we named this model *Initial AI*.

Additionally, to select samples for our experiment, we invited 30 undergraduate and master students to participate in our pilot test. One thousand test samples provided by the dataset are randomly divided into ten batches, and each batch is then randomly assigned to three students. These students were asked to label the main emotion type for each Weibo post and provide their confidence using a scale of 1 to 5. For the convenience of expression, we name these participants as *Pilot Human* in this study.

To improve the human-AI knowledge complementary to trigger the co-learning process, we only consider samples in which *Pilot Humans* have low confidence[3] but *Initial AI* have high confidence, or vice versa. Then for each participant, we will randomly select 32 tasks for the training phase, where both the initial AI and pilot humans have a 50% accuracy on the chosen samples[4]. We also follow the same strategy to select another 32 tasks for the testing phase. As we will report later, participants without AI support have a 51.2% average accuracy, which is not significantly different from that of the initial AI. This satisfies our goal of making the performance of AI and human comparable and moderate.

4.3 Condition Design

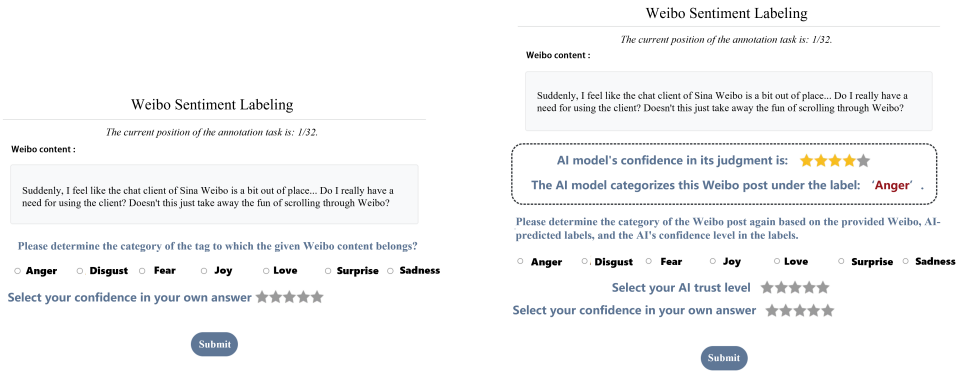
As our first study aims at identifying the human-AI co-learning process, we adopt a between-subject design and consider the below two conditions whereby their interface designs are reported in Figure 3.

- **Human Only.** In this condition, during the training phase, participants see Weibo posts, make judgments, and label their confidence without AI assistance or correctness feedback. Hence, no human-AI collaboration experience is involved.
- **AI Recommendation.** In this condition, during the training phase, the AI recommendation, accompanied by its confidence in the given recommendation, was provided when the participant was presented with the Weibo post.

4.4 Experiment Process Design

As shown in Figure 4, we mainly use a training-testing two-stage process design beyond the introduction and post-survey phases. In particular, upon arrival, participants first enter the *task introduction* interface, including brief information on the background.

[3] We chose the median of the confidence level as the cut-off point between high and low confidence levels for both Pilot Humans and Initial AI.
[4] Please refer to the support material Section A2 to check the sampling strategy.



(a) Human Only (b) AI Recommendation

Fig. 3: Experiment interface of the Study1.

Then, they were randomly assigned to one of the two conditions designed above and informed that they were about to enter an experiment, including the training, testing, and post-survey phases. For the participants who were assigned to the AI Recommendation condition, we also told them that during the training phase, they would be supported by an AI that outperformed the average human [72], but during the testing phase they had to complete the tasks without AI support. To motivate them to make accurate annotations carefully, they were also told that they should learn from the training phase, and beyond receiving a basic compensation, they will receive a substantial bonus according to their performance in the testing phase.

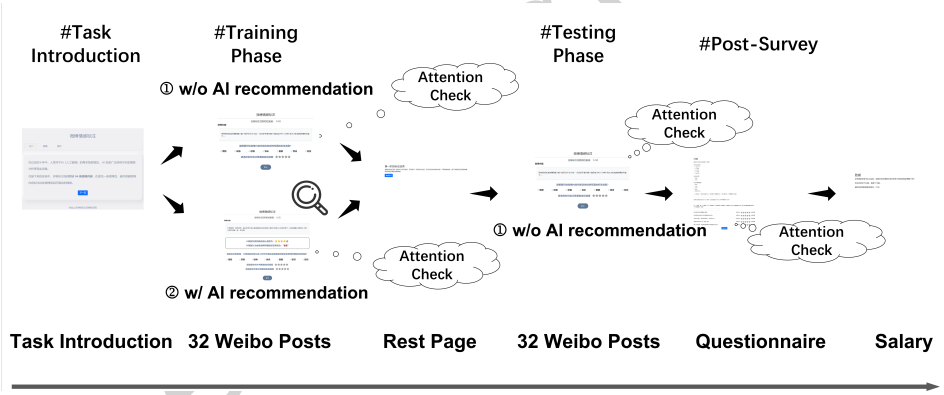


Fig. 4: Experiment Process Design. Please note that we have labeled the experimental interfaces in the figures corresponding to the training phase and testing phase. Specifically, the interface without AI recommendation is labeled as ①, which corresponds to Figure 3a, and the interface with AI recommendation is labeled as ②, which corresponds to Figure 3b.

After reviewing the task information, participants will enter the training phase to label the 32 Weibo posts selected from our sample pools using the sampling strategy described above. Note that the order of these samples was randomized to avoid the order effect. Afterward, participants will be taken to a page with prompts to enter the testing phase to complete another 32 tasks without AI support. They will further be asked to complete a post-survey that includes their demographic information, such as age, gender, and education.

In particular, we asked participants to take Frederick's cognitive reflection test (CRT) [31, 32, 73] so that we can identify the cognitive processes that an individual tends to engage with. As we used the three-questions version, the CRT score for each participant would be on a scale of 0 to 3. Additionally, for conditions involving AI, condition *AI recommendation* in study 1, we follow previous study [8, 10, 57, 74] to include three survey items to collect individuals' trust in AI, the perceived burden during task completion, and the level of engagement within the task using a scale of 1 to 5.

Finally, participants get bonuses based on their performance in the testing phase after finishing the post-survey. We also included attention-checking questions at each experiment phase so that we could filter out those inattentive participants. Only those who pass all three attention checks embedded in the training, testing, and post-survey phases will be considered valid.

4.5 Recruitment and Participant Demographics

We published our announcement on the WeChat Moments and a behavioral experiment recruitment Wechat group from the university to recruit participants for our study from October 12 to October 18, 2022[5]. All participants can only participate once. One hundred and eighty-two participants completed our first experiment. Twenty-four participants who did not pass our attention checks were excluded, leaving us with 158 participants[6]. The Mann-Whitney U-test reveals no significant difference for *age, education,* and *CRT score*, and a marginal difference ($p < 0.1$) for *gender* between these two conditions[7], suggesting participants were successfully randomly allocated to different experimental conditions.

4.6 Measures

To evaluate the co-learning process within the human-AI collaboration, we consider the following measures. First, for each participant, we collect their performances in the testing phase when they work alone without AI support and calculate:

- ***Human Accuracy***: the average accuracy scaling from 0 to 1 and a higher accuracy indicates a better performance.

[5] Our study was reviewed and exempted by our Institutional Review Board (IRB). Despite this, we still followed the same ethics and privacy requirements that an IRB would typically enforce. We acknowledge that this recruitment strategy may bring sample bias. However, it provides us with a special cultural and demographic context. We will discuss this limitation later.

[6] The detailed participant demographics for all three experiments are provided in the support material.

[7] To verify the significant difference between two samples, we use a Student t-test or Mann-Whitney U-test depending on whether the data follows the normal distribution.

- ***Human Confidence***: the average confidence scaling from 0 to 5 on the 32 testing tasks. Note that existing studies in human-AI collaboration highlight the necessity of people's overconfidence calibration [32, 72]. Given the improvement in human accuracy, the confidence reduction could reveal a confidence calibrating process that effectively reduces people's overconfidence in their own annotations [72]. In other words, a reduction in human confidence indicates a better learning outcome.

Furthermore, we use each participant's annotations in the training phase to finetune the initial AI, resulting in a *Finetuned AI*. Then we use the *Finetuned AI* to predict the labeling for the 32 tasks in the testing phase. Hence, we can calculate the performance, including *AI Accuracy, AI Confidence* for each finetuned AI in the testing phase. Note that as the finetuned AI is trained using participants' input, each participant will get a personalized finetuned AI. For each finetuned AI, we calculate

- ***AI Accuracy***: the average accuracy scaling from 0 to 1 and a higher accuracy indicates a better performance.
- ***AI Confidence***: the average confidence scaling from 0 to 5 on the 32 testing tasks. Unlike humans, a higher confidence for AI indicates a better performance in reaching model convergence.

Additionally, if the co-learning process occurs, human and AI are anticipated to exhibit greater similarity after their collaboration. Hence, we computed the alignment between human and AI annotations in the testing phase. And the correct alignment captures the beneficial aspects of their similarity and incorrect alignment captures the opposite side. For the *Human Only* condition, we assessed the alignment between human and the *Initial AI*, while for the *AI Recommendation* condition, we compared it with the *Finetune AI*.

- ***Alignment***: the percentage of tasks where participants and AI label the same and a higher alignment indicates a closer human-AI mental model.
- ***Correct/Incorrect Alignment***: the percentage of tasks where participants and AI both have correct or incorrect labeling. A higher value for correct alignment indicates a positive alignment for human-AI model, whereas a higher value of incorrect alignment indicates a negative alignment.

4.7 Results

For easy reading, we highlight the results into "Findings" and then report the empirical observations, including the p-value (p) and its effect size (r) if the result is significant.

Finding 1.1: Human can learn from the human-AI collaboration demonstrated as their accuracy improvement and confidence reduction when they work alone without AI support. As shown in Figure 5a and Figure 5b, participants who experienced AI recommendation during the training phase has a significantly higher ($p < 0.05, r = 0.141$) human accuracy than those in the *Human only* condition, indicating that participants benefit from the human-AI collaboration and improve their accuracy when working alone after their collaboration. Additionally, withdrawing AI support in the testing phase makes participants significantly less confident ($p < 0.05, r = 0.167$) in their annotations. Note that the confidence reduction in our

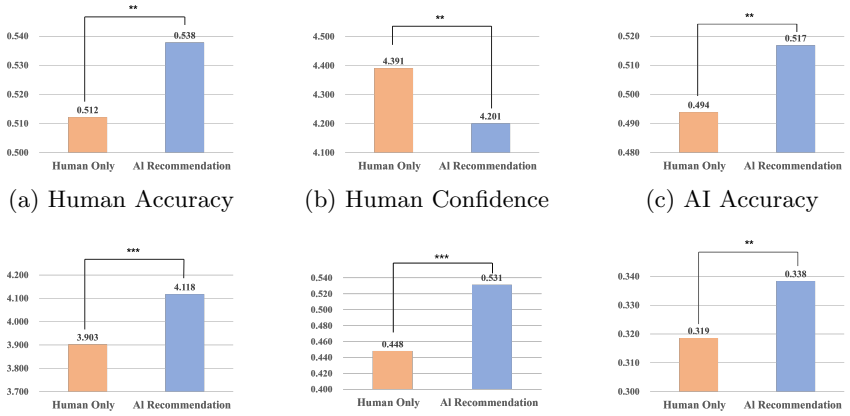


Fig. 5: The co-learning effect for the human-AI collaboration. $* : p < 0.1, ** : p < 0.05, *** : p < 0.01$.

study context indicates an effective confidence calibrating process [72]. Hence, **H1a is supported.**

Finding 1.2: AIs benefit from annotations generated through the human-AI collaboration, demonstrated as the significantly higher accuracy and confidence improvement after the fine-tuning process. For each participant in both conditions, we use his/her annotations from the training phase to train the initial AI into the fine-tuned AI, representing AI's learning from human. Note that each participant generates different annotations in the training phase so that the fine-tuned AI will be different for each participant. As shown in Figure 5c and Figure 5d, the AIs fine-tuned by the annotations generated from *AI Recommendation* condition demonstrate a significantly higher accuracy ($p < 0.05, r = 0.134$) and confidence ($p < 0.01, r = 1.069$) in the testing phase. This means that the human-AI collaboration helps participants generate better inputs to teach AIs to achieve better performance. Hence, **H1b is supported.**

Finding 1.3: The human-AI collaboration makes human and AI more aligned with each other, especially with more consistently correct annotations, even after their collaboration is no longer available. If the human-AI co-learning process occurs, we would expect their mental models to become closer and they should become more aligned. As shown in Figure 5e and Figure 5f, the human-AI collaboration enables participants (in the *AI recommendation* condition) and AIs (fine-tuned by the annotations generated from the *AI recommendation* condition) to achieve a significant higher alignment ($p < 0.01, r = 0.898$). Additionally, the human-AI collaboration led human and AI towards a significantly higher proportion of correct alignment ($p < 0.01, r = 0.744$). This implies that the human-AI co-learning process enables both human and AI to learn from each other through collaborations, and their mental models become closer in a positive way[8]. Hence, **H1c is supported.**

[8] We did not observe a significant incorrect alignment difference.

In summary, our study demonstrates that in our study context, the human-AI collaboration experience indeed triggers an effective co-learning process in both pathways that human can learn from the collaboration experience, AI can learn from the human's annotation, and they become more aligned with each other in a positive direction.

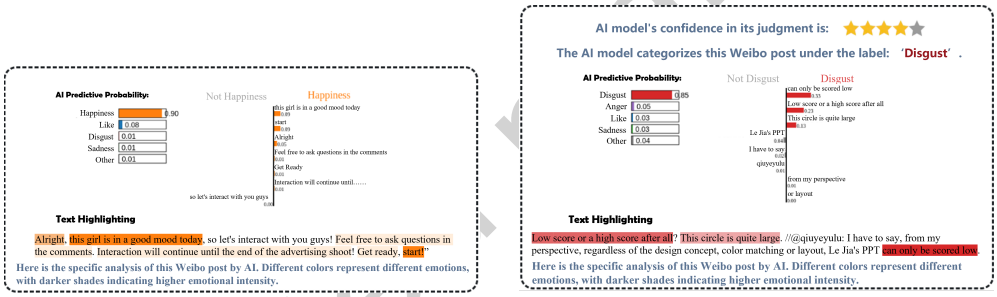
5 Study 2: The Vary Impacts of AI Feedback Designs

Given the demonstrated effective human-AI co-learning process in the first study, we design the second study to investigate *whether providing more AI feedback during the human-AI collaboration can enhance the co-learning process* to answer **RQ2**.

5.1 Conditions and Participants

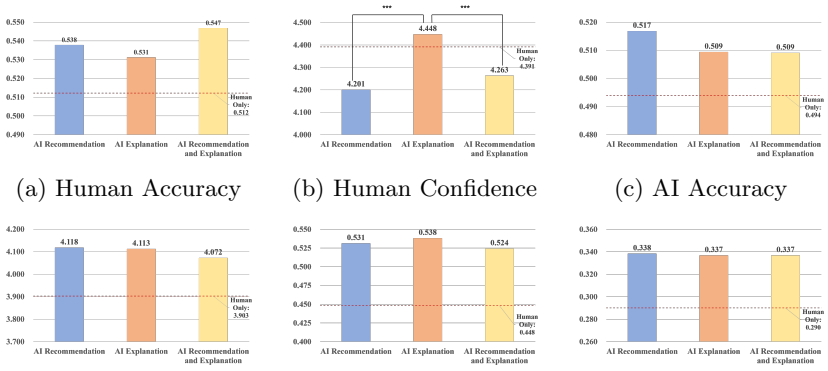
The AI feedback design, mainly focusing on providing AI recommendation or explanation, can affect the human-AI collaboration (i.e., [38, 59]). Hence, using the same task design as above, we consider the following three conditions whereby the interface designs are reported in Figure 6:

- ***AI Recommendation***: Just like the AI recommendation condition in Study 1.
- ***AI Explanation***: Participants only received explanations generated based on LIME (Local Interpretable Model Agnostic Explanation) without AI recommendation [75].
- ***AI Recommendation and Explanation***: The AI recommendation is accompanied by its confidence and the LIME-generated explanation simultaneously.



(a) AI Explanation (b) AI Recommendation and Explanation

Fig. 6: Experiment interface of the AI feedback. Note that the AI feedback within the dashed box varies under different conditions. Under the *AI Recommendation and Explanation* condition, the dashed box in Figure 3b displays the content shown in Figure 6b, while under the *AI Explanation* condition, it appears as in Figure 6a.



(a) Human Accuracy (b) Human Confidence (c) AI Accuracy

(d) AI Confidence (e) Alignment (f) Correct Alignment

Fig. 7: The AI feedback designs' impact on the co-learning effect. $* : p < 0.1, ** : p < 0.05, *** : p < 0.01$. The red dot line is the average value of the given measure in the *Human Only* condition from Study 1, representing the case where AIs and humans did not experience human-AI collaboration in the training phase.

We use the same protocol to recruit participants for these two additional conditions[9]. We have 211 participants in total for this Study. Additionally, the Kruskal-Wallis test shows no significant difference existing in *CRT score* among the three conditions above.

5.2 Results

To determine whether there are significant differences in the co-learning performance across the three AI feedback conditions, similar to previous studies [17, 32], we conducted a Kruskal-Wallis test. Then we would conduct pairwise comparisons using Dunn's tests with Benjamini-Hochberg adjustment if detecting a significant difference.

Finding 2.1: While different AI feedback designs have no significant differences in enhancing the co-learning performance regarding human accuracy, providing AI explanation makes participants more confident when they work alone in the testing phase. According to Figure 7a, human accuracy is not significantly different among the three conditions. However, under *AI Explanation* condition (Figure 7b), participants exhibited significantly higher confidence ($p < 0.01$) than the other two conditions. In other words, participants who only receive the AI explanation during the collaboration may exhibit greater confidence in their annotations when they work independently afterward, but their performance sees no significant improvement. One reason in play could be that when only an AI explanation is provided, participants will digest the explanation and confirm their own judgment, resulting in more overconfidence in the testing phase. **Hence, H2a is only partially supported but contrary to our hypothesis.**

[9] Note that we did not repeat condition *AI Recommendation* in Study 2 but reuse the data from Study 1. Participants' demographics are summarized in support material

Finding 2.2: Additional explanations during human-AI collaboration fail to trigger more AI learning or make their mental models closer. Figure 7c to 7f show no significant difference in both AI performance and human-AI consistency for AI fine-tuned by annotations from different AI feedback conditions. This suggests that providing more feedback during the human-AI collaboration does not contribute to AI learning or a more consistent mental model for human and AI. **Hence, H2b nor H2c is not supported.**

In summary, our study demonstrates that in our study context, where human and AI have an initial comparable and moderate performance, the AI feedback with explanations fails to trigger positive, if not negative, human-AI co-learning except by making humans more confident on their own.

6 Study 3: The Impact of Collaborative Workflow Designs

The third experiment aimed to answer **RQ3** by exploring *whether different collaborative workflows in human-AI collaboration would enhance human-AI co-learning process* when human and AI both have a moderate and comparable performance.

6.1 Conditions and Participants

Prior studies [17, 30, 48] provide mixed evidence on how a two-step collaborative workflow, which allows participants to update their initial decisions after being presented with the AI recommendation, can affect the human-AI collaboration. To investigate how such a workflow design could affect the human-AI co-learning process, using the same task design as before, we consider the following conditions:

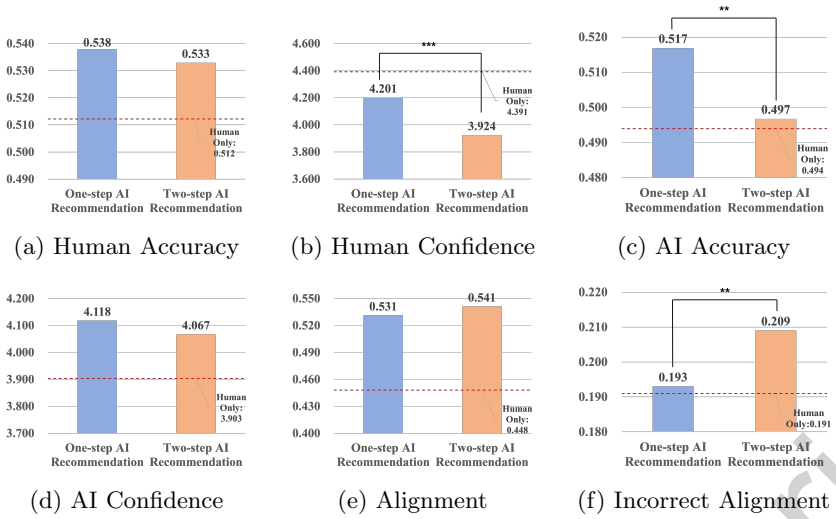
- **One-step AI Recommendation.** Just like the *AI Recommendation* condition in Study 1.
- **Two-step AI Recommendation.** In this condition, participants will read a post, make a judgment, and then receive both the AI's prediction and confidence level. At this point, they will make another judgment on the same post. They could change their answer but did not have to. This allows participants to double-think their previous annotation and update it, if they want to, during the training phase.

We use the same protocol to recruit participants to join the additional condition[10]. 133 participants are involved in this study. Additionally, there also exist no significant differences in *CRT score* between these two conditions.

6.2 Results

We compared participants' learning performance between one-step and two-step workflow conditions during the testing phase to investigate the impact of the collaborative workflow on the human-AI co-learning process.

[10] Again, we did not repeat condition *One-step AI Recommendation* but reuse the data from Study 1.



(a) Human Accuracy (b) Human Confidence (c) AI Accuracy

(d) AI Confidence (e) Alignment (f) Incorrect Alignment

Fig. 8: The collaborative workflow design's effect on the human-AI co-learning. $*$: $p < 0.1, **$: $p < 0.05, ***$: $p < 0.01$. Note that the red dot line is the average value of the given measure in the *Human Only* condition from Study 1, representing the case where no human-AI collaboration is involved in the training phase. Additionally, instead of reporting the insignificant correct alignment change, we focus on the results of incorrect alignment in this study.

Finding 3.1: The two-step workflow providing double thinking fails to yield stronger learning effects for human accuracy but further significantly reduces human confidence when they work alone afterward. The accuracy under different workflow configurations did not exhibit significant discrepancies (Figure 8a). However, when withdrawing the AI support in the testing phase, participants in the *two-step workflow* exhibit significantly lower confidence ($p < 0.01, r = 0.160$) (Figure 8b). Hence, the two-step collaborative workflow causes more significant reductions in human confidence when AI support is no longer available, indicating a stronger overconfidence calibration effect. Hence, **H3a is partially supported.**

Finding 3.2: For AIs, the two-step workflow design reduces the valuable information provided by human-AI collaboration. As shown in Figure 8c, the *two-step workflow* condition shows a significantly lower accuracy ($p < 0.05, r = 0.309$) for the fine-tuned AI than the *one-step workflow* condition. Additionally, according to Figure 8d, the AI's confidence also decreases, although not significantly. Hence, the double thinking provided by the *two-step workflow* configuration actually causes a negative impact on the "AI learning from human" process. Hence, **H3b is partially supported but in an opposite direction.**

Finding 3.3: Double-thinking provided by the two-step workflow design does not make human and AI more positively aligned with each other,

demonstrated as an insignificant change in correct alignment but a significant increase in incorrect alignment. When investigating the human-AI alignment in the testing phase, we observe an improvement in the alignment, although not significant according to Figure 8e. However, as shown in Figure 8f, when we further dive into the *incorrect alignment*, where human and AI have the same but incorrect annotation in the testing phase, we found a significant increase for the *two-step workflow* design ($p < 0.05, r = 0.285$). This indicates that the double-thinking provided by the *two-step workflow* design actually results in negative, rather than positive, consistency in annotations. Hence, **H3c is partially supported but in an opposite direction.**

In summary, the two-step collaborative workflow during the human-AI collaboration fails to yield stronger co-learning effectively and equally. It can more effectively calibrate human overconfidence for "human learning from AI", but make AI worse through the "AI learning from human" process. Moreover, although it slightly increases the human-AI mental model consistency, such an improvement in alignment is due to the significant incorrect alignment. Overall, these observations indicate an intervention-generated inequality effect of our collaborative workflow designs on the human-AI co-learning process.

7 Internal Audit: The Intervention-Generated Inequalities within Human-AI Co-Learning Process

As emphasized by existing studies [17, 35, 36], a design intervention, which refers to the AI feedback and collaborative workflow design in our study, may result in intervention-generated inequalities. An internal audit to disaggregate the observations is necessary. Hence, using the data collected from Study 2 and Study 3, we further take an internal audit to investigate how human cognitive reflection capability may result in an unequal learning process for the human-AI collaboration to answer **RQ4**.

7.1 Variables

Straightforwardly, we employed human accuracy (HumanAcc), human confidence (HumanConf), AI accuracy (AIAcc), AI confidence (AIConf), alignment (Align), correct alignment (CorrAlign) and incorrect alignment (InCorrAlign) during the testing phase as dependent variables, as they measure the co-learning outcome triggered by the human-AI collaboration. Additionally, as we aim to investigate the impact of AI feedback and workflow design, we consider them as independent variables and convert them into categorical variables. In particular,

- ***FEEDBACK***: we code the different AI feedback designs based on the information provided where *only provided AI Recommendation* is coded as 1, *only provide AI Explanation* is coded as 2 and *provide both AI Recommendation and Explanation* is coded as 3.

- ***WORKFLOW***: we label the different collaborative workflow designs according to whether the double-thinking mechanism is provided where *using one-step workflow* is coded as 1 and *using two-step workflow* is coded as 2.

Moreover, the learning outcomes in the testing phase are closely related to their performance during the training phase, as the outcome from training phase can reveal the effectiveness of the collaborative experience between human and AI. Hence, for each participant from Study 2 and Study 3 who experienced the human-AI collaboration, we evaluate the collaborative experience using their collaborative performance and alignment during the training phase:

- ***ColAcc***: the average accuracy from human-AI collaboration scaling from 0 to 1 in the training phase.
- ***ColAlign***: the percentage of tasks where participants and AI label the same in the training phase through the human-AI collaboration.

To control factors that might influence the co-learning outcome, we consider following self-reported metrics including individuals' trust in AI [10, 57], the perceived burden during the experiment [8, 57], and the level of engagement within the task using the average of three survey items we collected from the post-survey [74].

- ***Trust***: The average score from three post-survey questions assessing trust in AI, scaled from 1 to 5.
- ***Burden***: The average score from three post-survey questions assessing task burden, scaled from 1 to 5.
- ***Engagement***: The average score from three post-survey questions assessing task engagement, scaled from 1 to 5.

We also incorporated the average time taken to complete each annotation during the testing phase as a control variable.

- ***PageTime***: the average time spent on each page during the testing phase.

Finally, with the goal of uncovering potential intervention-generated inequalities arising from human cognitive reflection capabilities within the human-AI co-learning process and beyond considering all the participants from Study 2 and Study 3 as a whole[11], following the previous studies (i.e. [17, 35]), we divided our participants into two groups, where the **High_CRT** group has a CRT score of 3 and the **Low_CRT** group has a CRT score lower than 3[12].

7.2 Method

The descriptive statistics and correlations for all variables are reported in the support materials[13]. We calculated variation inflation factors (VIFs), and each was well below

[11] We report the result when considering all participants in Support materials and the results confirm our observations from Study 2 and Study 5, suggesting the robustness of our conclusion.
[12] We use this strategy as 57.08% of our participants have a CRT score of 3 and such a strategy can avoid the potential sample imbalance impact.
[13] The data utilized in regression model were subjected to Z-score normalization to eliminate the influence of dimensional discrepancies, with the exception of FEEDBACK and WORKFLOW.

the threshold of 10, with an average of 1.18 [76]. The condition number on the specification was 11.05, which is well below the threshold of 15 [77]. Hence, our results are not biased by multicollinearity.

7.3 Results

Table 1: Regression Results for Participants with High CRT Score

	HumanAcc	HumanConf	AIAcc	AIConf	Align	CorrAlign	IncorrAlign
FEEDBACK	0.079	0.199*	-0.199*	-0.160*	-0.174	-0.143	-0.109
	(0.109)	(0.109)	(0.109)	(0.082)	(0.108)	(0.107)	(0.111)
WORKFLOW	0.024	-0.459**	-0.344	-0.522***	-0.294	-0.347	-0.068
	(0.222)	(0.222)	(0.222)	(0.166)	(0.22)	(0.218)	(0.225)
ColAcc	0.033	0.012	0.055	0.056	0.180**	0.163*	0.073
	(0.088)	(0.088)	(0.088)	(0.066)	(0.088)	(0.087)	(0.089)
ColAlign	0.187*	-0.227**	-0.137	0.620***	0.234**	0.135	0.222**
	(0.11)	(0.11)	(0.11)	(0.082)	(0.109)	(0.108)	(0.111)
Pagetime	-0.098	0.028	-0.147*	-0.018	-0.164**	-0.169**	-0.067
	(0.079)	(0.079)	(0.079)	(0.059)	(0.078)	(0.078)	(0.08)
Trust	0.007	0.223**	-0.032	0.049	0.066	0.027	0.061
	(0.09)	(0.09)	(0.09)	(0.068)	(0.089)	(0.089)	(0.091)
Burden	-0.024	-0.021	0.081	0.043	-0.057	-0.018	-0.067
	(0.08)	(0.08)	(0.08)	(0.06)	(0.079)	(0.078)	(0.081)
Engagement	0.017	0.262***	-0.063	-0.041	0.05	0.019	0.047
	(0.078)	(0.079)	(0.078)	(0.059)	(0.078)	(0.077)	(0.079)
_cons	-0.006	-0.091	0.429***	0.266**	0.282**	0.325**	0.068
	(0.138)	(0.138)	(0.138)	(0.103)	(0.137)	(0.136)	(0.14)
N	157	157	157	157	157	157	157
r2	0.043	0.200	0.070	0.372	0.127	0.086	0.067

Table 2: Regression Results for Participants with Low CRT Score

	HumanAcc	HumanConf	AIAcc	AIConf	Align	CorrAlign	IncorrAlign
FEEDBACK	0.064	-0.089	0.149	-0.007	0.142	0.192	0.004
	(0.116)	(0.119)	(0.125)	(0.086)	(0.127)	(0.125)	(0.138)
WORKFLOW	-0.108	-0.495*	-0.201	-0.078	0.481*	0.162	0.528*
	(0.253)	(0.26)	(0.273)	(0.188)	(0.276)	(0.272)	(0.3)
ColAcc	0.151*	-0.01	0.081	0.06	0.055	0.129	-0.062
	(0.087)	(0.09)	(0.094)	(0.065)	(0.095)	(0.094)	(0.104)
ColAlign	0.105	-0.031	-0.125	0.695***	0.259**	0.174*	0.196*
	(0.094)	(0.096)	(0.101)	(0.07)	(0.102)	(0.101)	(0.111)
Pagetime	-0.147*	-0.228**	0.026	0.002	0.01	-0.091	0.124
	(0.089)	(0.091)	(0.096)	(0.066)	(0.097)	(0.095)	(0.105)
Trust	0.094	0.106	0.057	0.122*	-0.041	0.043	-0.101
	(0.095)	(0.098)	(0.103)	(0.071)	(0.104)	(0.103)	(0.113)
Burden	-0.217**	-0.028	0.035	-0.057	-0.109	-0.154	0.01
	(0.089)	(0.092)	(0.097)	(0.066)	(0.098)	(0.096)	(0.106)
Engagement	0.071	0.089	0.136	-0.187***	0.002	0.07	-0.078
	(0.092)	(0.095)	(0.099)	(0.068)	(0.1)	(0.099)	(0.109)
_cons	-0.048	0.079	-0.085	0.001	-0.246	-0.245	-0.094
	(0.154)	(0.159)	(0.166)	(0.114)	(0.168)	(0.166)	(0.183)
N	111	111	111	111	111	111	111
r2	0.119	0.113	0.076	0.588	0.118	0.096	0.097

As reported in Table 1 and Table 2, by dividing our participants based on their CRT score, we can observe that:

For the *human learns from AI* subprocess, the AI feedback designs demonstrate a significant positive relation with human confidence for participants with a high CRT score. However, this effect is not significant for participants with a low CRT score. In fact, we observe a negative although not significant coefficient, which is markedly

different from participants with a high CRT scores ($p = 0.05$). Moreover, the workflow design demonstrates a significant negative relation on human confidence for both groups, while the coefficient for those with a low cognitive capability is larger than for those with a high cognitive capability. This means that the overconfidence calibration effect for participants with high reflection capability is weaker for participants with low reflection capability. Hence, **H4a is partially supported but in an opposite direction.**

For the *AI learns from human* subprocess, we observe a more significant negative impact from AI feedback and collaborative workflow for participants with a high CRT score, especially a significant negative impact in AI accuracy. Hence, the double-thinking mechanism provided through a two-step workflow design can be harmful to AI learning, especially for those with a high CRT score. Hence, **H4b is partially supported but in an opposite direction.**

For the *human-AI alignment*, we observe an significant opposite influence when compared the regression coefficients of FEEDBACK and WORKFLOW between groups with different cognitive reflection capabilities. Moreover, for participants with low CRT capability, the workflow design is significantly positive to the alignment, especially to the incorrect alignment. This means that compared with participants with high CRT score, double-thinking provided by the workflow design make those with low CRT score more overreliance to the AI's response. Hence, **H4c is partially supported but in an opposite direction.**

Overall, the internal audit analysis confirms an intervention-generated inequality effect within the co-learning process for participants with different cognitive reflection capability. More over, the negative effect is more significant for participants with high cognitive reflection capability, while the feedback and workflow design can result into a more significant overreliance effect for participants with low cognitive reflection capability.

8 Discussions, Implementations and Limitations

8.1 Answering Questions: The Human-AI Co-Learning Process

Can human-AI collaboration trigger an effective co-learning process? Clearly, the human-AI collaboration experience during the training phase triggers an effective human-AI co-learning process, resulting in performance improvement when they work alone in the testing phase: humans have an accuracy improvement, and their overconfidence is calibrated when working alone. AIs fine-tuned by the annotations generated through human-AI collaboration have higher accuracy and confidence. Humans and AIs experiencing the human-AI collaboration also have a closer mental model demonstrated as both a higher alignment and correct alignment. Beyond some previous studies (i.e. [17, 24–27]) reporting no or limited learning phenomenon, our study demonstrates the effective learning process and confirms the importance of collaborative experience for both pathways of the human-AI co-learning process, which is consistent with the learning cycle theory.

How can different AI feedback designs influence the co-learning process? Generally speaking, there exist no significant differences in co-learning among the

three AI feedback designs, except that providing only AI explanation worsens human overconfidence, which is contrary to the expectation. One reason for these observations is that unlike existing study [17], which provides the ground truth as the AI recommendation and provides the prepared explanation, in our experiments, human and AI have comparable and moderate accuracy. The additional explanation, while demonstrating the AI's confidence and reasoning, also emphasizes its uncertainty and the potential inconsistencies with human decision logic. Notably, given the lack of a ground truth, meaning users were unaware of whether their or the AI's answers were correct, the immediate benefits of the AI's explanation were unclear but imposing additional cognitive burdens. This duality may negatively impact users' perceptions of the AI, ultimately diminishing their acceptance and trust in AI's recommendation [78]. This led humans to entirely ignore the explanations and enhanced their overconfidence [43, 60]. In other words, learning from the AI feedback makes humans distrust AI and disrupts the co-learning process. Hence, similar to existing discussions on explanations' potential negative toward human-AI collaboration [3, 4, 37, 41, 43, 60], the AI feedback based on explanations need to be carefully designed, considering its complexity, user's expectation and faithfulness, to enhance the learning process.

How can different collaborative workflow designs influence the co-learning process? We can see that the two-step collaborative workflow design in human-AI collaboration demonstrates an overconfidence calibration effect for human, which is consistent with existing studies focusing on the decision-support setting [17, 30, 48]. However, it results in negative effects on AI's learning performance from the *AI learning from human* process. Additionally, while we observe a slightly increase alignment, such an improvement is due to the increase in incorrect alignment, suggesting an overreliance effect, which is contrary to the expectation of introducing such a cognitive forcing strategy [17, 30, 32]. One reason for this phenomenon is that in our study, where humans and AIs both have a moderate performance with an accuracy rate of approximately 50%, the collaborative workflow design further reduces participants' confidence in their own annotations. This is because the double-thinking mechanism necessitates a re-evaluation of analyses that have already been completed, imposing additional cognitive costs [79] and enlarging their self-doubt on their answer in the first round. Consequently, this extra cognitive burden diminishes participants' ability to identify the AI's wrong answer. Rather, they intend to follow AI's recommendation, which leads to the increasing overreliance. When AI is further finetuned with these outputs with wrong data, its accuracy and confidence become worse.

How does the cognitive reflection capability in human-AI collaboration affect the dual-pathway co-learning process? We confirm that cognitive reflection capability (CRT) disproportionately affects human and AI's learning process within the same collaborative configuration. This is consistent with existing studies investigating the intervention-generated inequalities effect in decision support setting [17, 32, 37]. More importantly, by distinguishing the two pathways of the human-AI co-learning process, we found that the AI's learning process being more significantly impacted. Moreover, humans with high CRT exhibit characteristics of overconfidence, while those with low CRT tend to overly reliant on their AI partners. The design of AI feedback and collaborative workflow further amplifies these differences. In other

words, contrary to existing studies arguing that the intervention designs, which is the AI feedback and collaborative workflow design in this study, can benefit people with high cognitive reflection capability more [17, 32, 37], they can also lead to more negative impact. One reason of these observations can be that participants with high cognitive reflection capabilities, who more frequently engage with conscious processing ("System 2"), may exhibit a more analytical disposition towards digesting the provided information [31]. Consequently, they are better equipped to comprehend and interpret AI feedback, meaning they will engage more consciously with AI during collaboration. Digesting the inconsistent AI feedback, especially the explanations, leads them to rely less on AI [80] and more on their own superior analytical skills to complete tasks independently, thereby building up overconfidence. Conversely, the double-thinking mechanisms, which can balance intuition ("System 1") and analytical ("System 2") thinking, help calibrate this overconfidence [17, 48]. Hence, the generated annotations from their collaboration will be too inconsistent with AI's previous mental model so that finetuning with these data would then reduce AI's performance. However, for participants with low cognitive reflection capability, they intends to follow AI's recommendation, especially when providing the double-thinking mechanisms. Hence, the generated annotations from the collaboration fail to provide more valuable information to improve AI's performance, except improving their alignment.

8.2 Theoretical Contribution

Our research primarily contributes to two domains: human-AI collaboration and learning cycle theory.

Firstly, while most existing studies have concentrated on human-AI collaboration within an AI-supported decision-making context [10, 18], this study considers AI as a partner [6, 8, 18, 50] rather than merely a tool [3, 10], and extends the scope of research into the co-earning mechanism. Similar to existing studies that have reported promising effects of human-AI collaboration designs on improving collaborative outcomes, this study demonstrates that such experiences in human-AI collaboration can improve individual post-collaboration performance, suggesting an effective co-learning process and validating the efficacy of this design. However, while many existing studies reveal the effectiveness of AI feedback [29] and collaborative workflow [11, 17, 30, 48, 81] design to enhance human's cognitive engagement with AI information to promote human-AI collaboration, we further substantiate that the co-learning process does not occur spontaneously [17, 58] and show that these designs can also harm the co-learning process, resulting into human's self-overconfidence, overreliance to AI and AI's performance reduction, which is consistent with recent studies re-evaluating the AI feedback and collaborative workflow designs' effectiveness on human-AI collaborations [3, 4, 17, 30, 43, 48]. Additionally, similar to existing studies (i.e., [4, 32, 37]), our study also confirms the intervention-generated inequality due to the difference in cognitive reflection capability. However, contrary to our expectation, participants with high cognitive reflection capability performed worse in the co-learning process, and this intervention-generated inequality is also unequal for the two pathways in the co-learning process. In other words, the benefits of human-AI collaboration designs

for people with high reflection capability can also be harmful. This indicates a potential mediation effect between the high reflection capability and the benefit from the human-AI collaboration designs, which is worth further exploring.

Secondly, for learning cycle theory, while existing research has primarily explored the team learning dynamics [82, 83] and interaction within human teams [21, 84], we incorporate AI as a partner to explore the factors that affect their learning process within such a hybrid human-AI collaboration context, enriching the limited understanding of the factors that stimulate and influence the learning process [22]. While existing research has established that this learning cycle does not occur spontaneously but is influenced by numerous factors [14, 21, 22], we have further explored how designs of human-AI collaboration and cognitive reflection capabilities impact the learning cycle. Specifically, we distinguish two pathways and find that the human-AI collaboration designs' impact on the learning outcome varies within the two pathways while the participants' cognitive reflection capability further amplifies this difference. Moreover, our study also finds that the human-AI collaboration designs aiming at improving the learning process can turn out to be negative for the learning outcome. Hence, paying attention to the potential negative impacts and mitigating them when designing the learning environment with AI is necessary, which aligns with existing design principles [17, 32, 37].

8.3 Design Implication

This empirical evidence highlights the design principles that we could follow to facilitate the co-learning process embedded in human-AI collaboration.

"Something is better than nothing." While existing studies have reported mixed observations on whether human-AI collaboration can achieve complementary performance, resulting in a lack of unified design principles [2, 3, 30, 47], this study focuses on situations where human and AI performance are comparable, affirming the necessity of human-AI collaboration. Our study provides strong evidence that in the context where human and AI have comparable and moderate performance (around 50% accuracy), human and AI can learn from their collaboration and gain better performance when they work alone. Hence, beyond the potential complementary performance in an AI-assisted human decision-making setting, the human-AI collaboration, even if they both only have moderate performance, can trigger the co-learning process and improve each other. In other words, whenever applicable, setting up human-AI collaboration could be beneficial for both parties. For example, introducing an AI partner in educational or skill training settings to create a collaborative learning experience [21, 23, 28], even when both human and AI only have a moderate performance, can still trigger a co-learning process to improve both human and AI.

"More can be less." Existing literature acknowledges the critical nature of AI feedback and collaborative workflow designs for human-AI collaboration [4, 10, 18, 32]. While it suggests that increased feedback or deeper thinking can be beneficial, our findings indicate that mindless accumulation of designs may result in negative effects. The insignificant differences from additional AI feedback in Study 2 and the two-step collaborative workflow in Study 3 suggest that making the interaction between human and AI more complex is not helpful. Rather, the significantly negative effects

we observe suggest that naively deploying a complex AI system without optimizing for their collaboration could lead to negative impacts on co-learning. Given existing research demonstrating that people are more likely to trust AI when AI exhibits high performance [10, 85], it is suggested that different settings of co-learning should adopt various configurations of collaboration rather than assuming that more complexity is better. When both human and AI performance are moderate, a simple collaborative configuration should initially be deployed to trigger the co-learning process. Once both parties have significantly improved through co-learning, explanations or collaborative workflows can subsequently be incorporated to ensure that AI advice receives due consideration in the decision-making process. For instance, in scenarios involving human-AI collaboration within cybersecurity operations, it may be beneficial to provide humans with the judgments and explanations only when the AI model has high confidence. Conversely, for judgments where AI's confidence is low, only the AI's judgment might be presented to avoid unnecessary cognitive burden and potential misleading.

"One size can not fit all." Existing research highlights that differing cognitive capabilities can lead to intervention-generated inequalities when designing human-AI collaboration mechanisms [17, 32, 37], our study takes a step further to highlight the intervention inequality, including both AI feedback and collaborative workflow designs, on the two pathways of human-AI co-learning for participants with different cognitive capabilities. This suggests that considering the participants' cognitive reflection capability [4, 32, 37], to construct an adaptive learning environment [33, 34], could be helpful when deploying an AI system. Taking AI-assisted clinic training as an example, we can provide high CRT users with more detailed explanations and double-thinking mechanisms to challenge their potential overconfidence tendencies. For low CRT users, given their intention to follow AI's recommendation, we can design a guided decision-making process to encourage them to conduct independent analysis while considering AI recommendations, especially when AI's recommendation is correct. Through such an adaptive learning design, we can better balance the needs of users with different cognitive reflection capabilities, thereby optimizing the effectiveness of the human-AI co-learning process.

8.4 Limitations and Further Work

Our current study has several limitations, which open venues for future studies. First, we focus on a scenario where human and AI have comparable and medium accuracy to enlarge their knowledge of complementary regions. However, it is valuable to assess scenarios with different accuracy levels.

Secondly, we opted for the emotion annotation task as our single-task scenario. It is worth considering tasks in more consequential and objective scenarios, like medicare training or cybersecurity operations.

Another important limitation is participant recruitment. We recruited participants from WeChat Moments and a behavioral experiment recruitment group, rather than the traditional crowdsourcing platform like MTurk, LabintheWild, or Prolific [17], which could be explored in future studies under different socioeconomic contexts.

Furthermore, our study is an online experiment using ChineseMacBert, a pre-trained natural language model. This can provide inspiration for investigating the human-AI collaboration with the rising large language models (LLMs) like ChatGPT, GPT-4, etc. Moreover, we randomly selected 32 samples in the "training phase" and "testing phase", which already took an average of 22.17 minutes for participants to finish the experiment. We acknowledge that it may take more than thirty-two samples to make their mental models converge. Therefore, further research using more advanced models and improving the sample selection strategy could be a promising direction to explore, and incorporating more features of reinforcement learning in interactive learning scenarios should also be considered.

Finally, we take a two-step, between-subject design for this study. However, we did not investigate the long-term aspect, such as the iteration of the co-learning process following an active-learning paradigm, which is worth further exploration. While we successfully unfold the influences from AI feedback, collaborative workflow, and human cognitive capability on the co-learning process, exploring the intrinsic mechanisms of how these designs intertwine with each other to impact the co-learning outcome is still in its early stage. This could be highly valuable for further in-depth studies.

9 Conclusion

This study investigates the human-AI co-learning process within their collaboration, by distinguishing the dual-pathway learning process, which includes both human and AI learn from each other. By conducting three online, two-stage, between-subject behavioral experiments, we provide empirical evidence for the effective co-learning process in human-AI collaboration, leading to improved performance during the subsequent independent work in the testing phase. However, we found that the AI feedback design, the collaborative workflow, and human cognitive capability can intertwine with each other, demonstrating the unequal and even negative impacts on the two pathways. These observations suggest the necessary but non-trivial of effective human-AI collaboration designs to trigger both pathways in the co-learning process to achieve complementary performance.

Hence, we anticipate further research to explore mechanisms that result in unequal effects in the two learning pathways, which will allow for a more refined optimization of AI deployment to strengthen human-AI collaboration and mitigate its potential negative effects.

References

[1] Capel, T., Brereton, M.: What is Human-Centered about Human-Centered AI? A Map of the Research Landscape. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), p. 22 (2023). https://doi.org/10.1145/3544548.3580959

[2] Bansal, G., Nushi, B., Kamar, E., Weld, D.S., Lasecki, W.S., Horvitz, E.: Updates in human-ai teams: Understanding and addressing the performance/compatibility

tradeoff. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 2429–2437 (2019)

[3] Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M.T., Weld, D.: Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1–16 (2021)

[4] Zhang, Y., Liao, Q.V., Bellamy, R.K.: Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 295–305 (2020)

[5] Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P.N., Inkpen, K., *et al.*: Guidelines for human-ai interaction. In: Proceedings of the 2019 Chi Conference on Human Factors in Computing Systems, pp. 1–13 (2019)

[6] Liang, C., Proft, J., Andersen, E., Knepper, R.A.: Implicit communication of actionable information in human-ai teams. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–13 (2019)

[7] Pinski, M., Adam, M., Benlian, A.: Improving Human-AI Collaboration with AI Knowledge Effects of a Human Enablement for Efficient AI Delegation. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), p. 17 (2023). https://doi.org/10.1145/3544548.3580794

[8] Bansal, G., Nushi, B., Kamar, E., Lasecki, W.S., Weld, D.S., Horvitz, E.: Beyond accuracy: The role of mental models in human-ai team performance. In: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, vol. 7, pp. 2–11 (2019)

[9] Siemon, D.: Elaborating team roles for artificial intelligence-based teammates in human-ai collaboration. Group Decision and Negotiation **31**(5), 871–912 (2022)

[10] Yin, M., Wortman Vaughan, J., Wallach, H.: Understanding the effect of accuracy on trust in machine learning models. In: Proceedings of the 2019 Chi Conference on Human Factors in Computing Systems, pp. 1–12 (2019)

[11] Ma, S., Lei, Y., Wang, X., Zheng, C., Shi, C., Yin, M., Ma, X.: Who should i trust: Ai or myself? leveraging human and ai correctness likelihood to promote appropriate trust in ai-assisted decision-making. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, pp. 1–19 (2023)

[12] Huang, K., Zhou, J., Chen, S.: Being a solo endeavor or team worker in crowdsourcing contests? it is a long-term decision you need to make. Proceedings of the ACM on Human-Computer Interaction **6**(CSCW2), 1–32 (2022)

[13] Gimpel, H., Graf-Seyfried, V., Laubacher, R., Meindl, O.: Towards artificial intelligence augmenting facilitation: Ai affordances in macro-task crowdsourcing. Group Decision and Negotiation **32**(1), 75–124 (2023)

[14] Koole, B.: Trusting to learn and learning to trust. A framework for analyzing the interactions of trust and learning in arrangements dedicated to instigating social change. Technological Forecasting and Social Change **161**(August 2019), 120260 (2020) https://doi.org/10.1016/j.techfore.2020.120260

[15] Wenskovitch, J., North, C., Tech, V.: Intelligence : Designing for the Two Black Boxes Problem. IEEE Computer **53**(8), 29–39 (2020)

[16] Das, D., Chernova, S.: Leveraging rationales to improve human task performance. In: Proceedings of the 25th International Conference on Intelligent User Interfaces, pp. 510–518 (2020)

[17] Gajos, K.Z., Mamykina, L.: Do People Engage Cognitively with AI? Impact of AI Assistance on Incidental Learning. In: International Conference on Intelligent User Interfaces, Proceedings IUI, pp. 794–806 (2022). https://doi.org/10.1145/3490099.3511138

[18] Yu, K., Berkovsky, S., Taib, R., Zhou, J., Chen, F.: Do i trust my machine teammate? an investigation from perception to decision. In: Proceedings of the 24th International Conference on Intelligent User Interfaces, pp. 460–468 (2019)

[19] Kolb, D.A.: Management and the learning process. California management review **18**(3), 21–31 (1976)

[20] Tsai, M.-T., Lee, K.-W.: A study of knowledge internalization: from the perspective of learning cycle theory. Journal of knowledge management **10**(3), 57–71 (2006)

[21] Morris, T.H.: Experiential learning–a systematic review and revision of kolb's model. Interactive learning environments **28**(8), 1064–1077 (2020)

[22] Lee, K.-W., Tseng, Y.-F.: Driving the dual learning process of management knowledge: A social cognitive theory perspective. The International Journal of Management Education **22**(1), 100940 (2024)

[23] Kolb, D.: Experiential learning: experience as the source of learning and development pearson education. Inc. https://doi. org/10.1017/CBO9781107415324 **4** (2015)

[24] Xu, C., Lien, K.C., Höllerer, T.: Comparing Zealous and Restrained AI Recommendations in a Real-World Human-AI Collaboration Task. In: Conference on Human Factors in Computing Systems - Proceedings (2023). https://doi.org/10.1145/3544548.3581282

[25] Lee, Y.-c., Yamashita, N., Huang, Y.: Exploring the Effects of Incorporating Human Experts to Deliver Journaling Guidance through a Chatbot **5**(April), 1–27 (2021)

[26] Tsai, C.-H., You, Y., Gui, X., Kou, Y., Carroll, J.M.: Exploring and promoting diagnostic transparency and explainability in online symptom checkers. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1–17 (2021)

[27] Melsión, G.I., Torre, I., Vidal, Eva and Leite, I.: Using Explainability to Help Children Understand Gender Bias in AI. In: Interaction Design and Children, pp. 87–99 (2021). https://doi.org/10.1145/3459990.3460719

[28] Gibbons, S., Ebbeck, V., Gruno, J., Battey, G.: Impact of adventure-based approaches on the self-conceptions of middle school physical education students. Journal of Experiential Education **41**(2), 220–232 (2018)

[29] Kim, S.S.Y., Watkins, E.A., Russakovsky, O., Fong, R., Monroy-Hernández, A.: "Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), p. 17 (2023). https://doi.org/10.1145/3544548.3581001

[30] Green, B., Chen, Y.: The principles and limits of algorithm-in-the-loop decision making. Proceedings of the ACM on Human-Computer Interaction **3**(CSCW), 1–24 (2019)

[31] Frederick, S.: Cognitive reflection and decision making. Journal of Economic perspectives **19**(4), 25–42 (2005)

[32] Duan, X., Ho, C.J., Yin, M.: The Influences of Task Design on Crowdsourced Judgement: A Case Study of Recidivism Risk Evaluation. WWW 2022 - Proceedings of the ACM Web Conference 2022, 1685–1696 (2022) https://doi.org/10.1145/3485447.3512239

[33] Hwang, G.-J., Sung, H.-Y., Chang, S.-C., Huang, X.-C.: A fuzzy expert system-based adaptive learning approach to improving students' learning performances by considering affective and cognitive factors. Computers and Education: Artificial Intelligence **1**, 100003 (2020)

[34] Yaghmaie, M., Bahreininejad, A.: A context-aware adaptive learning system using agents. Expert Systems with Applications **38**(4), 3280–3286 (2011)

[35] Veinot, T.C., Mitchell, H., Ancker, J.S.: Good intentions are not enough: how informatics interventions can worsen inequality. Journal of the American Medical Informatics Association **25**(8), 1080–1088 (2018)

[36] Bertel, S., Dressel, T., Kohlberg, T., Jan, V.: Spatial knowledge acquired from pedestrian urban navigation systems. In: Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services, pp. 1–6 (2017)

[37] Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: An overview of interpretability of machine learning. In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), pp. 80–89 (2018). IEEE

[38] Ghai, B., Liao, Q.V., Zhang, Y., Bellamy, R., Mueller, K.: Explainable active learning (xal): An empirical study of how local explanations impact annotator experience. arXiv preprint arXiv:2001.09219 (2020)

[39] Wang, X., Yin, M.: Watch Out for Updates: Understanding the Effects of Model Explanation Updates in AI-Assisted Decision Making. In: Conference on Human Factors in Computing Systems - Proceedings, p. 19 (2023). https://doi.org/10.1145/3544548.3581366

[40] Poursabzi-Sangdeh, F., Goldstein, D.G., Hofman, J.M.: Manipulating and Measuring Model Interpretability, (2021). https://doi.org/10.1145/3411764.3445315

[41] Eiband, M., Buschek, D., Kremer, A., Hussmann, H.: The impact of placebic explanations on trust in intelligent systems. In: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–6 (2019)

[42] Papenmeier, A., Kern, D., Englebienne, G., Seifert, C.: It's complicated: The relationship between user trust, model accuracy and explanations in ai. ACM Transactions on Computer-Human Interaction (TOCHI) **29**(4), 1–33 (2022)

[43] Smith-Renner, A., Fan, R., Birchfield, M., Wu, T., Boyd-Graber, J., Weld, D.S., Findlater, L.: No explainability without accountability: An empirical study of explanations and feedback in interactive ml. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–13 (2020)

[44] Hassenzahl, M., Borchers, J., Boll, S., Pütten, A.R.-v.d., Wulf, V.: Otherware: How to best interact with autonomous systems. Interactions **28**(1), 54–57 (2020)

[45] Cila, N.: Designing Human-Agent Collaborations : Commitment , responsiveness , and support. In: CHI Conference on Human Factors in Computing Systems, pp. 1–18 (2022)

[46] Aydoğan, R., Baarslag, T., Gerding, E.: Artificial intelligence techniques for conflict resolution. Group Decision and Negotiation **30**(4), 879–883 (2021)

[47] Holstein, K., McLaren, B.M., Aleven, V.: Student learning benefits of a mixed-reality teacher awareness tool in ai-enhanced classrooms. In: International

Conference on Artificial Intelligence in Education, pp. 154–168 (2018). Springer

[48] Buçinca, Z., Malaya, M.B., Gajos, K.Z.: To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. Proceedings of the ACM on Human-Computer Interaction **5**(CSCW1) (2021) https://doi.org/10.1145/3449287 arXiv:2102.09692

[49] Jung, M.F., Šabanović, S., Eyssel, F., Fraune, M.: Robots in groups and teams. In: Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, pp. 401–407 (2017)

[50] Wildman, J.L., Nguyen, D., Thayer, A.L., Robbins-Roth, V.T., Carroll, M., Carmody, K., Ficke, C., Akib, M., Addis, A.: Trust in human-agent teams: A multilevel perspective and future research agenda. Organizational Psychology Review, 20413866241253278 (2024)

[51] Larasati, R., De Liddo, A., Motta, E.: Meaningful explanation effect on user's trust in an ai medical system: Designing explanations for non-expert users. ACM Transactions on Interactive Intelligent Systems **13**(4), 1–39 (2023)

[52] Chiang, C.-W., Lu, Z., Li, Z., Yin, M.: Are two heads better than one in ai-assisted decision making? comparing the behavior and performance of groups and individuals in human-ai collaborative recidivism risk assessment. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, pp. 1–18 (2023)

[53] Zhang, R., Flathmann, C., Musick, G., Schelble, B., McNeese, N.J., Knijnenburg, B., Duan, W.: I know this looks bad, but i can explain: Understanding when ai should explain actions in human-ai teams. ACM Transactions on Interactive Intelligent Systems **14**(1), 1–23 (2024)

[54] Kim, J., Cho, Y.H.: My teammate is ai: understanding students' perceptions of student-ai collaboration in drawing tasks. Asia Pacific Journal of Education, 1–15 (2023)

[55] Bosch, K., Schoonderwoerd, T., Blankendaal, R., Neerincx, M.: Six Challenges for Human-AI Co-learning. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **11597 LNCS**, 572–589 (2019) https://doi.org/10.1007/978-3-030-22341-0_45

[56] Holstein, K., Aleven, V., Rummel, N.: A conceptual framework for human–AI hybrid adaptivity in education. In: Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 12163 LNAI, pp. 240–254 (2020). https://doi.org/10.1007/978-3-030-52237-7_20

[57] Zagalsky, A., Te'Eni, D., Yahav, I., Schwartz, D.G., Silverman, G., Cohen, D.,

Mann, Y., Lewinsky, D.: The Design of Reciprocal Learning between Human and Artificial Intelligence. Proceedings of the ACM on Human-Computer Interaction **5**(CSCW2) (2021) https://doi.org/10.1145/3479587

[58] Schoonderwoerd, T.A.J., Zoelen, E.M.v., Bosch, K., Neerincx, M.A.: Design patterns for human-AI co-learning: A wizard-of-Oz evaluation in an urban-search-and-rescue task. International Journal of Human Computer Studies **164**(February), 102831 (2022) https://doi.org/10.1016/j.ijhcs.2022.102831

[59] Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., *et al.*: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information fusion **58**, 82–115 (2020)

[60] Kulesza, T., Stumpf, S., Burnett, M., Yang, S., Kwan, I., Wong, W.-K.: Too much, too little, or just right? ways explanations impact end users' mental models. In: 2013 IEEE Symposium on Visual Languages and Human Centric Computing, pp. 3–10 (2013). IEEE

[61] Lim, B.Y., Dey, A.K., Avrahami, D.: Why and why not explanations improve the intelligibility of context-aware intelligent systems. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 2119–2128 (2009)

[62] Czerwinski, M., Cutrell, E., Horvitz, E.: Instant messaging and interruption: Influence of task type on performance. In: OZCHI 2000 Conference Proceedings, vol. 356, pp. 361–367 (2000)

[63] Bialek, M., Pennycook, G.: The cognitive reflection test is robust to multiple exposures. Behavior research methods **50**, 1953–1959 (2018)

[64] Blair, D.J.: Experiential learning for teacher professional development at historic sites. Journal of Experiential Education **39**(2), 130–144 (2016)

[65] Murphy, L., Wilson, J., Greenberg, S.: Equine-assisted experiential learning in occupational therapy education. Journal of Experiential Education **40**(4), 366–376 (2017)

[66] Rosenthal, S.L., Dey, A.K.: Towards maximizing the accuracy of human-labeled sensor data. In: Proceedings of the 15th International Conference on Intelligent User Interfaces, pp. 259–268 (2010)

[67] Fede, J.H., Gorman, K.S., Cimini, M.E.: Student employment as a model for experiential learning. Journal of Experiential Education **41**(1), 107–124 (2018)

[68] Isaak, J., Devine, M., Gervich, C., Gottschall, R.: Are we experienced? reflections on the suny experiential learning mandate. Journal of Experiential Education **41**(1), 23–38 (2018)

[69] Stanovich, K.E., West, R.F.: Advancing the rationality debate. Behavioral and brain sciences **23**(5), 701–717 (2000)

[70] Gao, J., Yang, Z., Li, F., Yasen, B., Wen, S.: The influence of cognitive ability in chinese reading comprehension: can working memory updating change chinese primary school students' reading comprehension performance? Frontiers in Psychology **14**, 1283781 (2023)

[71] Cui, Y., Che, W., Liu, T., Qin, B., Wang, S., Hu, G.: Revisiting Pre-trained Models for Chinese Natural Language Processing. In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 657–668 (2020)

[72] Hou, Y.T.Y., Jung, M.F.: Who is the Expert? Reconciling Algorithm Aversion and Algorithm Appreciation in AI-Supported Decision Making. Proceedings of the ACM on Human-Computer Interaction **5**(CSCW2) (2021) https://doi.org/10.1145/3479864

[73] Sinayev, A., Peters, E.: Cognitive reflection vs. calculation in decision making. Frontiers in psychology **6**, 532 (2015)

[74] Greene, B.A.: Measuring Cognitive Engagement With Self-Report Scales: Reflections From Over 20 Years of Research. Educational Psychologist **50**(1), 14–30 (2015) https://doi.org/10.1080/00461520.2014.989230

[75] Ribeiro, M., Singh, S., Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144 (2016). https://doi.org/10.18653/v1/n16-3020

[76] O'brien, R.M.: A caution regarding rules of thumb for variance inflation factors. Quality & quantity **41**, 673–690 (2007)

[77] Belsley, D.A., Kuh, E., Welsch, R.E.: Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. John Wiley & Sons, Boston (2005)

[78] Jiang, Y., Nagarajan, V., Baek, C., Kolter, J.Z.: Assessing generalization of sgd via disagreement. arXiv preprint arXiv:2106.13799 (2021)

[79] Fogliato, R., Chappidi, S., Lungren, M., Fisher, P., Wilson, D., Fitzke, M., Parkinson, M., Horvitz, E., Inkpen, K., Nushi, B.: Who goes first? influences of human-ai workflow on decision making in clinical imaging. In: 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 1362–1374 (2022)

[80] Wang, X., Lu, Z., Yin, M.: Will you accept the ai recommendation? predicting human behavior in ai-assisted decision making. In: Proceedings of the ACM Web Conference 2022, pp. 1697–1708 (2022)

[81] Fogliato, R., Chouldechova, A., Lipton, Z.: The Impact of Algorithmic Risk Assessments on Human Predictions and its Analysis via Crowdsourcing Studies. Proceedings of the ACM on Human-Computer Interaction **5**(CSCW2) (2021) https://doi.org/10.1145/3479572 arXiv:2109.01443

[82] Kim, R.H., Gilbert, T., Ristig, K., Chu, Q.D.: Surgical resident learning styles: faculty and resident accuracy at identification of preferences and impact on absite scores. Journal of Surgical Research **184**(1), 31–36 (2013)

[83] Borredon, L., Deffayet, S., Baker, A.C., Kolb, D.: Enhancing deep learning: Lessons from the introduction of learning teams in management education in france. Journal of Management Education **35**(3), 324–350 (2011)

[84] Sims, R.R.: Kolb's experiential learning theory: A framework for assessing person-job interaction. Academy of Management Review **8**(3), 501–508 (1983)

[85] Wiczorek, R., Meyer, J.: Effects of Trust , Self-Confidence , and Feedback on the Use of Decision Automation. Frontiers in Psychology **10**(March), 519–112 (2019) https://doi.org/10.3389/fpsyg.2019.00519