Interaction of Thoughts: Towards Mediating Task Assignment in Human-AI Cooperation with a Capability-Aware Shared Mental Model

Ziyao He*
marsroscope@stu.xjtu.edu.cn
Xi'an Jiaotong University
MOE KLINNS Lab
Xi'an, Shaanxi, China

Yunpeng Song*
yunpengs@xjtu.edu.cn
Xi'an Jiaotong University
MOE KLINNS Lab
Xi'an, Shaanxi, China

Shurui Zhou
shuruiz@ece.utoronto.ca
University of Toronto
Toronto, Ontario, Canada

Zhongmin Cai†
zmcai@sei.xjtu.edu.cn
Xi'an Jiaotong University
MOE KLINNS Lab
Xi'an, Shaanxi, China

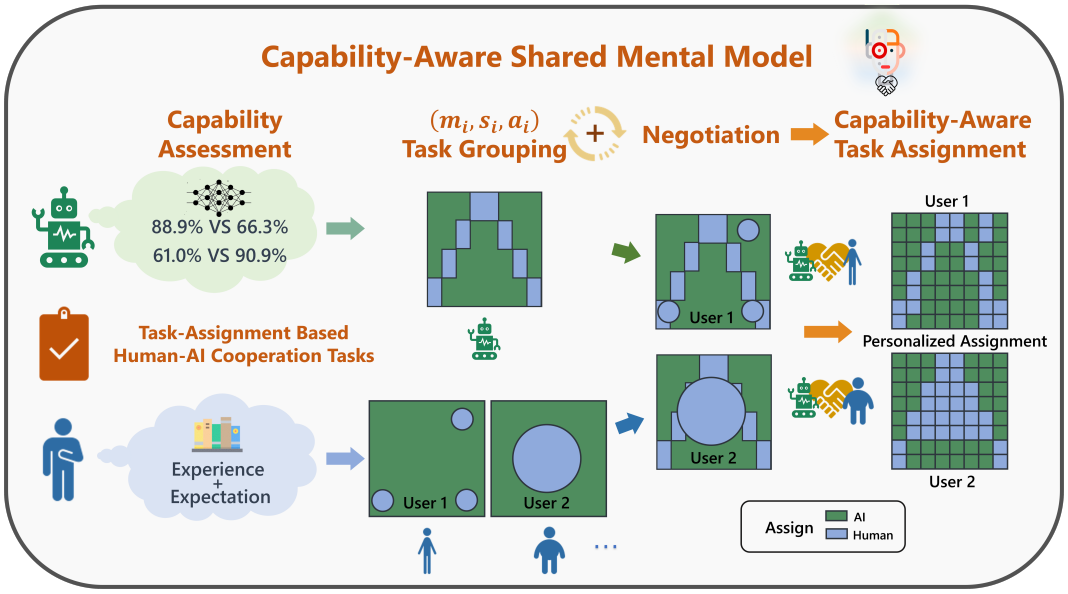

Figure 1: The process of forming capability-aware shared mental model for task assignment of human-AI cooperation.

ABSTRACT

The existing work on task assignment of human-AI cooperation did not consider the differences between individual team members regarding their capabilities, leading to sub-optimal task completion results. In this work, we propose a capability-aware shared mental model (CASMM) with the components of task grouping and negotiation, which utilize tuples to break down tasks into sets of scenarios relating to difficulties and then dynamically merge the task grouping ideas raised by human and AI through negotiation. We implement a prototype system and a 3-phase user study for the proof of concept via an image labeling task. The result shows building CASMM boosts the accuracy and time efficiency significantly through forming the task assignment close to real capabilities within few iterations. It helps users better understand the capability of AI and themselves. Our method has the potential to generalize to other scenarios such as medical diagnoses and automatic driving in facilitating better human-AI cooperation.

*Contribute Equally to this work.
†Corresponding Author

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04…$15.00
https://doi.org/10.1145/3544548.3580983

CCS CONCEPTS

• **Human-centered computing → HCI theory, concepts and models**; **User studies**; *Interaction design theory, concepts and paradigms.*

KEYWORDS

shared mental model, human-AI cooperation, task assignment

ACM Reference Format:

Ziyao He, Yunpeng Song, Shurui Zhou, and Zhongmin Cai. 2023. Interaction of Thoughts: Towards Mediating Task Assignment in Human-AI Cooperation with a Capability-Aware Shared Mental Model. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3544548.3580983

1 INTRODUCTION

Artificial intelligence (AI) technology is permeating all sectors of society, and AI has evolved into an intelligent supporter capable of performing various tasks. However, for real-world tasks, high accuracy with low fault tolerance is required. Since AI cannot guarantee 100% accuracy, a more practical and feasible way to use the power of AI for increasing efficiency and reducing human labor is to pair humans and AI together to complete these tasks. In fact, some interesting pioneering works have been carried out in this direction, such as automated driving [62], medical diagnosis [4], loan approval [60], employee recruitment [42], and pair programming [64]. In various forms of human-AI pairing, a basic form of cooperation is to send tasks to AI or human depending on the AI and human capabilities, so as to optimize certain performance measures of the human-AI team, such as the accuracy of task completion and efficiency. In this way, it is rational to make AI only work on cases in which it is good at, while asking human to deal with those complex and unexpected cases that require flexibility or out-of-domain knowledge. We term this form of AI-human pairing or teaming as Task Assignment based Human-AI cooperation (TAHAC).

Many tasks in human-AI cooperation can be directly modelled as TAHAC such as autonomous driving with human supervisory or collaborative disease diagnosis on medical images. The key challenge is to determine which task should be assigned to which team member, i.e., to human or to AI, so that humans and AI could collaboratively achieve high accuracy as well as decent time efficiency in completing a series of real-world tasks. For example, in a scenario where human doctors team up with AI assistants to diagnose diseases such as detecting breast cancer metastases in images of lymph node tissue sections [45], human-AI teams need to decide which cases need to be consulted with human doctors, and which could be completed by an AI assistant with high confidence. False diagnosis can delay patient treatment, while manual diagnosis is often time-consuming. Letting AI handle tasks where a confirmed correct diagnosis can be obtained quickly by the AI, and reassigning the rest to human doctors will increase the time efficiency of the human-AI team while ensuring a higher accuracy rate.

In real-world scenarios, it is difficult to achieve effective task assignments since it has to overcome several challenges. Firstly, the assignment of tasks should be based on the capabilities of each team member. However, the lack of understanding of mutual capabilities between humans and AI may further lead to mutual distrust or over trust, leading to failed cooperation. This is often due to the fact that human collaborators have unrealistic expectations of AI, i.e, human's estimations of AI's capabilities are not calibrated to what AI's actual capabilities are in a timely manner [7]. For example, a driver completely ignored manual supervision while driving in bright light, leading to a fatal crash that cost his life [61]. And research has shown that it is because the driver did not realize the inability of the automated driving algorithm to distinguish between a white sky and the roof of a speeding white truck in a bright light scenario [59]. Secondly, task allocation should take into account not only the differences in capabilities between humans and AI, but also the differences between various human individuals. The differences in human capabilities largely affect their behaviors when they cooperate with AI. For example, Glick et al. [24]'s research showed that novice clinicians tend to rely on automated diagnostic algorithms, while experts are prone to perform manual diagnoses. Hence, it may be appropriate to assign human a difficult task when AI and human experts cooperate. But when the human counterpart changes to become a novice, asking the AI to complete the task can become a better choice. In such scenarios, methods as Wilder et al. [83] proposed, which utilize the average performance distribution of the human to train a unified task assignment algorithm for different humans may not be the best practice. Consequently, to solve these challenges, the assignment of tasks for human-AI cooperation should not only be based on the mutual understanding of human and AI's capabilities to avoid distrust or over trust, but also need to consider the different capabilities of different members to dynamically produce the assignment plan for different human-AI teams.

To this end, we propose to form a **capability-aware shared mental model (CASMM)** that takes into account capabilities between humans and AI, which is able to represent human and AI's capabilities and dynamically mediate their work assignments during the cooperation process. The *shared mental model* is a notion that was first studied in human-human cooperation [29, 37]. It ensures that every team member gets hold of each other's mental model, thus reaching an agreement on the team's behavior. Here, inspired by works in systematic engineering [29] and human-robot cooperation [34, 53], a mental model could be defined as a mental representation or explanation that an agent (human or AI) uses to interact, describe, and predict the behavior of a particular substance or system [78]. Additionally, during the procedure of forming the CASMM, the two parties share not only their completion results but also the reliability evaluations for each other (such as accuracy or confidence) of completing cooperational tasks, which is shown to be effective in cognitive science for cooperation between humans [2].

To formally represent the team members' capabilities of conducting certain tasks, we utilize a tuple (m_i, s_i, a_i), in which m_i denotes the team member (human or AI), s_i denotes the type of task in a specific scenario (e.g., driving in fog, in bright sunlight, or at rainy night), and a_i represents the performance of the team member handling such tasks. The tuple distinguishes the same task in different scenarios (*task grouping*), serves as the basis for building a shared mental model, and clearly shows the mapping between specific tasks and the capability of each team member to perform the task. Additionally, the process of building a shared mental model

often needs multiple rounds of interaction to help humans and AI to better understand each other's capabilities in order to make a better decision on the task assignment. Specifically, we designed *negotiation* methods to dynamically merge the task grouping ideas raised by human and AI. Such negotiations are multiple-round, in order to catch up with the latest grouping as the members gain a deeper understanding of each other. As cooperation progresses, if either team member believes the task assignment does not reflect their real capabilities, our method will update the value of corresponding tuples, thus the shared mental model will be revised as well to better represent each team member's capability. As a result, for task assignment between human-AI teams, **our shared mental model is modified as a collaboratively negotiated and dynamically constructed model for grouping tasks based on each team member's capability that could be further mapped into task assignment.** With the support of CASMM, for real-world cases such as medical diagnosing, the physician and the medical imaging AI could come up with capability-aware task assignments so that each member diagnoses the images it is better at. As mutual understanding gradually improves during cooperation, the shared mental model is continuously updated, leading towards better task assignments and more effective cooperation.

To evaluate the effectiveness and usefulness of the proposed CASMM, we implement a prototype system focused on the task of collaborative image labelling. Specifically, we designed a 3-phase user study, where human and AI cooperate on the tasks through a user interface. From the experiment, we observe that assigning tasks based on a shared mental model boosts the accuracy and efficiency of the human-AI team performance significantly. The results also show that building the shared mental model for human-AI task assignment is able to form a simple, intuitive task assignment plan capable of mediating complementary human-AI cooperation with relatively few iterations. The human participants also are more aware of AI's capabilities. The shared mental model is not task-specific and can be extended to more general scenarios such as loan approval, medical diagnoses, and automatic driving in facilitating better human-AI cooperation.

Our contributions of this work are summarized as follows:

1 We propose to utilize the notion of the shared mental model (SMM) to facilitate task assignment based human-AI cooperation (TAHAC), which is a collaboratively constructed model for grouping tasks based on human and AI's capabilities and dynamically mediating their work assignment during the cooperation process.
2 We propose a unified form of tuples (m_i, s_i, a_i) to represent the task-specific capability of human and AI, which is understandable for both human and AI as well as feasible for being computed, stored, negotiated, revised, and further used for assigning tasks. Utilizing the forms of tuples as the basis, we designed a mechanism to build capability-aware shared mental model (CASMM) via iterative multi-round task grouping and task assignment negotiation.
3 We implement a prototype system, conduct a 3-stage user study to evaluate the effectiveness of the proposed CASMM and explore the dynamics of human-AI cooperation. The results illustrate that the CASMM can improve the accuracy and time efficiency for task assignment based human-AI cooperation (TAHAC). Moreover, the CASMM can help human-AI teams better understand each member's capability and then come up with a task assignment plan to better fit their real capabilities.

2 RELATED WORK

2.1 Human-Robot Cooperation and Human-AI Cooperation

2.1.1 Human-robot cooperation

Nowadays, a significant portion of robots includes AI algorithms to accomplish tasks more autonomously and intelligently. As El Zaatari et al. [17] described, human–robot cooperation can be categorized into 4 kinds: independent, simultaneous, sequential, and supportive, according to the various degrees of interaction. As such categorizing suggests, the robots for human-robot cooperation tasks are often designed to accomplish tasks that are physically difficult or impossible for humans to accomplish, and lots of works usually default a commander-follower relationship between human and robot [34]. Hence, unlike the TAHAC we discussed, human-robot cooperation less often involves the issues we have mentioned about whether the current task should be assigned to humans or robots. However, a series of works done to mediate task assignments in human-robot cooperation are still enlightening for us. A series of works focus on sharing robots' explanations of their behavior with human [7, 85] or other robots [23], thus assisting human operators to come up with assignments more accord with the robots' capability. For example, Chakraborti et al. [7] evaluated various explanation generation algorithms under rescue scenarios that come up with explanations best fitting human expectations in order to make human better understand and operate the robots. Another research direction is trying to make robots mimic human experts' actions, thus making human forms natural assignments as cooperating with fellow human workers. Dehkordi et al. [14] leveraged human experts' series of actions or logic of cooperating for training, so robots are able to predict what might need to be done for human based on training. Although these researches partly rely on several robotic concepts or characteristics such as environmental awareness (or situational awareness) and the robots' capability of switching macro motion state, the way of mimicking human or sharing explanations of behaviors share similar motivation for us to build capability-aware shared mental models.

2.1.2 Human-AI cooperation

Nowadays most of the engineering tasks between human and AI adopt the method of human-AI cooperation, which includes complete and incomplete task assignments [43]. The former is similar to human-robot cooperation, human and AI complete different types of jobs. The later human works on the basis of AI's pre-work to improve the subsequent work or, as what we defined, the task assignment based human-AI cooperation (HATAC forms). In practice, HATAC forms exist in many fields, such as automated driving [62], medic diagnosis [4], loan approval [60], human recruiting [42] and pair programming [64]. For example, in the employee recruitment process, human recruitment managers would screen some of the pre-screened candidates by AI to efficiently process candidate data from a broader and more diverse talent pool [42].

Several methods are adopted to further boost HATAC tasks in an effective way. First of all, some works paid attention to designing information-delivering methods for human during the scenarios when AI agents fail or finish the task with low confidence [30]. For instance, Matthew Kay et al. studied visualization to help people better perceive the uncertainty of the output given by mobile predictive system [33]. Secondly, further, several studies are devoted to explaining AI for human-AI cooperation, thus making human capable of getting the various information of the AI they are cooperating with. These explanations cover perspectives from AI capability assessment [25, 50], intermediate output [25, 27], attribution analysis of reasoning process [1, 50, 71, 84] and algorithmic output [63, 82] etc. Among these, what motivates our designing of the shared mental model is that works like [66, 67] utilized textual knowledge and feature expressions to make AIs more causable for these explanations. For example, Schaekermann et al. [66] select knowledge from medical guides such as rules to classify segments of a polysomnogram to one of five sleep stages to explain the accordance of AI's judgment of the patient's sleep phase. Jean Y. Song et al. replaced the numerical amount of 3D training data with responses aggregated from human worker annotations while viewing different video frames of the same 3D object [77]. These works described above inspired us to design our shared mental model by explaining to humans in a unified form the cases in which AI is prone to failures that can allow humans to better understand the capabilities of AI.

Several works take another route of elaborating human factors (such as history logs and training data from humans) into calibrating machine learning algorithms [36] for completing human-AI cooperation tasks. Chancellor et al. [8] utilized the users' tags of photos on Instagram together with clinician annotations to predict varied mental illness severity online. Works done in crowd-sourcing incorporate and aggregate multiple human worker's outputs and corresponding history tracking, then try to decide whether the collection of human samples is enough to infer the correct answer with acceptable confidence, thus stop consulting human workers [26, 31, 38, 46, 54, 76, 83]. Wilder et al. [83] introduced human factors by designing cost and utility based on the average human-capability distribution for algorithms deciding when to query human workers in experiments. For such work, collecting enough human work data to form an average or distributed representation of capabilities is an important part of reaching its algorithmic goals. However, the same well-trained algorithm is not fully compatible with human collaborators with different capabilities. For example, there is a large difference between the trusting behavior of novice and expert physicians during their cooperation with diagnostic AI [24]. Moreover, the designing of such query algorithm requires parameters from the AI that participates in the cooperation tasks, which usually are inaccessible since many commercialized AI models are not open-source. Also, in this way, algorithms still remain opaque for human users, which also might undermine human's trust in AI. These works inspired us to incorporate different individuals' capabilities as factors into the task assignment process, thus leading to the idea of building a shared mental model.

2.2 Mental Models in Teams

As we described, to form proper work assignment of human-AI hybrid intelligent teams, both human and AI agents need to form a decent mental model of each other [40] and themselves [15, 49].

Mental models and shared mental models for human teams: mental model was originally conceptualized by Craik (1943), as a model of thinking that parallels reality [12]. This notion was further developed and re-clarified in various domains for effective utilization [57]. The mental model shared in teams contains how team members predict what the others are going to do, hence it facilitates coordinating actions between teammates [11, 29]. For human-human teams, usually, team members with similar mental models perform better than those with more accurate but less similar mental models [48]. Therefore, the main purpose of forming a shared mental model in human cooperation teams, such as sports teams, is to reach a shared knowledge state, in which the knowledge held by each member about the upcoming actions of the team is at least similar to other team members' knowledge of these actions [13, 16, 79]. Work as [51] described that a common shared mental model between humans should involve mutual behaviors like information sharing, group learning, and cognitive consensus among team members [51].

Mental models of human in human-AI cooperation: the mental model of human-workers significantly influences the efficiency of human-AI cooperation. Therefore, several works considered human mental models in evaluating human-AI cooperation [10, 41, 81]. They focus on decoding human perception or expectation of robots or AI into AI during the cooperation process [7, 14, 23]. Zhang et al. [88] composed a case study illustrating various factors influencing how human workers accept AI assistants and expect their performances, which include pre-existing attitudes and past experiences. Several systems involving human-AI cooperation incorporate the notion of humans' mental models of AI assistants. For example, Bansal et al. [4] mentioned the notion of error boundary of AI agents perceived by human and discussed how properties of error boundaries affect the mental model of human workers and team efficiency. Gero et al. [22] apply the think-aloud method to study people's mental models of AI in a cooperative word guessing game and evaluate under which case people could get a better estimation of AI assistants in games. Nguyen et al. [52] designed interfaces to help humans build a more accurate mental model of AI by presenting information such as dynamic relations between AI's predicted correctness and sources for judgment in a fast-checking task. These designs try to help human in building a unidirectional mental model of AI while familiarizing with an AI system's algorithms. However, building only one-side mental model is insufficient, as Harmanpreet Kaur stated, a shared mental model needed to be formed among team members [32].

Shared mental models in human-AI cooperation: upon the purposes of bridging human's teaming up with AI, lots of works choose to explore whether and how the shared mental model would boost human-robot cooperation [14, 47, 53, 56–58, 68]. It is probably due to the fact that robots have more bionic properties and physical characteristics that human workers are able to easily anthropomorphize, rationalize, relate and compare. Several works

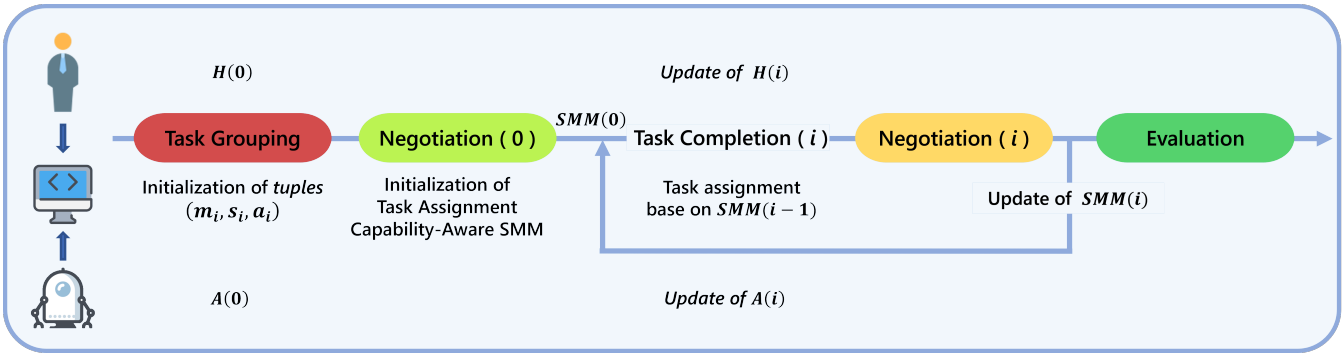

Figure 2: The iterative process of building capability-aware shared mental model between human and AI.

tried to identify and embed the shared mental model into the cooperation process. Such as Nikolaidis and Shah [53] proposed to utilize shared mental model as an influential factor of the Partially Observable Markov Decision Process (POMDP), Scheutz [68] broke down shared mental models into facts and beliefs, and then defined processes of updating tokens to iteratively update belief and propositions within the team.

Although several works have noticed the importance of incorporating shared mental models into more generalized human-AI teams, the explorations remain in treating mental models as an infecting factor, or researching whether the existence of such notion would boost teamwork under specific tasks [86]. For example, Razin et al. [62] evaluated Banks et al. [3]'s Human-AI shared mental model theory by examining how a self-driving vehicle's hazard assessment facilitates shared mental models. Claire et al. used the decision tree to simulate teammate's mental models and make sense of the hints in card games [44]. As in [32], currently, the bottleneck in human-AI cooperation is the socio-technical gap between the fluid intents and interactions of humans and the discrete and brittle features of AI agents, which will persist until both humans and AI can better understand each other's capabilities. However, as best as we could learn, no work has been done in generalized human-AI cooperation scenarios that highlight or incorporate a shared mental model for task assignment within the systems. Additionally, as we described in section 2.1.2, the current work did not take the capability difference between individuals into consideration. Therefore, in this study, we propose a method to derive a capability-aware shared mental model for task assignment in order to facilitate better cooperation between human and AI.

3 CAPABILITY-AWARE SHARED MENTAL MODEL IN TASK ASSIGNMENT BASED HUMAN AND AI COOPERATION (CASMM-TAHAC)

In this paper, we focus on mediating tasks where human and AI cooperate together to complete the same tasks and need to decide which portion of the tasks should be assigned to which member. A large portion of the conflicts in such task assignments for human-AI cooperation tasks is due to the team workers' inadequate understanding of each other's capability which leads to inappropriate assignment distribution. In fact, human has diverse capabilities and their understanding of AI's capability is gradually developing along the cooperation, which should be taken into consideration to influence the task assignment as well. To this end, we utilize the notion of the mental model and further develop the capability-aware shared mental model for task assignment in human-AI cooperation.

3.1 Designing the Shared Mental Model

As ideas raised in systematic engineering [29] and human-robot cooperation [34, 53], a mental model is usually the mental representation or explanation that an agent (human or AI) uses to interact with, describe, and predict the behavior of a certain matter or system. For TAHAC tasks, we define mental models as the mental representation formed by each side of the group (human and AI) consisting of the capability evaluations of each member and assignment strategies conditioned on the evaluations. This notion of "mental model" may seem different from the "standard" definition in cognitive science. However, it does reflect the ideas of how human or AI perceive the counterpart's capability and predict each other's behaviors in task assignment, which is in consistence with the mental model derived in the area of human-AI cooperation [34, 53]. Since the human part and AI part may form divergent mental models, we also further develop the shared mental model for TAHAC. Specifically, a shared mental model is referred to as the consistent mental model shared among the team members after reaching a consensus. Such consensus should base on a unified representation for both sides to align the capabilities, and rounds of negotiations to collaboratively eliminate the divergence and dynamically catch up with the updates of the mental models. It does not necessarily imply that team members share identical mental models, but hold compatible mental models that lead to common expectations for the task and team [74].

Given the above considerations, there exist several issues in front of building a shared mental model between human and AI for task assignment. First of all, different from the communication between humans, human and AI is not able to understand each other easily through a few conversations. Building a communication bridge between human and AI so that they can understand each other's capabilities is a difficult but essential problem. Secondly, once the human-AI team recognizes each other's capability, a clear and reasonable task division strategy should be reached between human

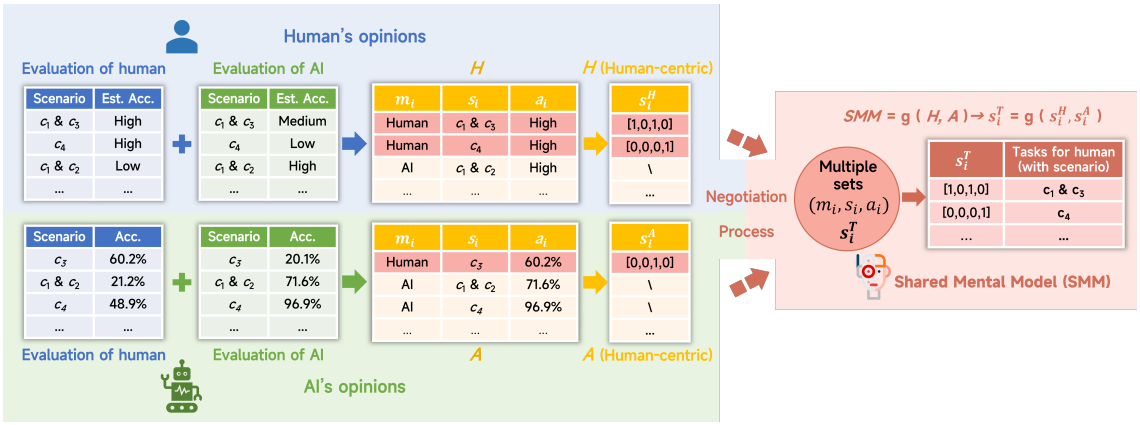

Figure 3: An example of the process of building the capability-aware shared mental model between human and AI. H represents the list of tuples from human's perspective (human mental model), and A represents the list of tuples from AI's perspective (AI Mental Model). H and A are simplified into 2 lists of s_i^H and s_i^A as human-centric representations, which denotes the scenarios to assign tasks to human from human or AI's perspective. SMM represents the shared mental model, which is computed by merging s_i^H and s_i^A into the list of s_i^T (T as for human-AI Team) through negotiation methods and then mapped into capability-aware task assignment.

and AI based on the individual capability to obtain the optimal reconciliation [18]. Moreover, effective evaluation methods and standards should be introduced to precisely guide the updates of the shared mental model in the cooperation procedure. To address the above issues, we propose to design a capability-aware shared mental model in task assignment based human-AI cooperation with the following components, i.e., task grouping and negotiation. The task grouping describes how well a certain member deals with the specific task scenario. The negotiation eliminates the divergences and reaches a consensus between the human-AI team regarding distributing the task scenarios according to each member's strengths and weaknesses. Hence, the process of building such shared mental model is illustrated as in Fig.2. In task grouping, based on the evaluation of each other's performance of a certain amount of tasks with ground truth, human and AI are first guided by the system to form each other's mental model $(H(0), A(0))$ in the form of tuples. Then, the negotiation process (negotiation (0)) merges the mental models of both parties to form an initialized shared mental model $(SMM(0))$ for the system to assign the upcoming cooperational tasks without ground truth. When human and AI cooperate to complete the assigned tasks (task completion (i)), they can enter negotiation sessions (negotiation (i)) several times to further update and revise their mental model $(H(i), A(i))$ based on the comparison of task assignment and their capability assessment, and then form the dynamically updated $SMM(i)$. After the completion of the TAHAC tasks, designed evaluation metrics can be utilized to assess the effectiveness of human-AI cooperation based on SMM, thus evaluating the effectiveness of our CASMM for TAHAC tasks.

3.2 Task Grouping

In a TAHAC situation, in order to assign members with the most suitable tasks according to their capabilities, we design the first step of building the shared mental model as **task grouping**, which groups the tasks into varied **scenarios** that human and AI exhibit significant differences in capabilities of completing tasks, and then builds the mapping between each member's capability and these scenarios utilizing a unified form of tuples (m_i, s_i, a_i).

Defining scenarios. Scenarios are the task groups that human and AI exhibit significant differences in capabilities of completing tasks. The intuition of grouping tasks into scenarios is based on two aspects. Firstly, humans are observed to tend to develop simple cases to understand the performance of complex systems [55]. Secondly, many existing studies have demonstrated that machine learning algorithms have been observed to have "biases" for some scenarios [19, 83] which are observable for human, despite the fact that AI and humans make judgments based on different criteria when completing tasks. Hence, such **scenarios**, in other words, are observable "features" (or attributes) that have impacts on the performances of machine learning algorithms from a certain aspect, as many of these scenarios are macro-observable and understandable for humans. For example, many image recognition algorithms are not good at recognizing blurred images, or when images' color and grayscale information is erased (with only edge or silhouette information present) [20]. For an autonomous driving vehicle based on machine learning image recognition algorithms, driving scenarios that produce blurred visuals (such as foggy or rainy days) or silhouette-only observations (night driving with low-light) may significantly affect the accuracy and efficiency of autonomous driving algorithms. Such observations, even without insights of machine learning algorithms, can still be observed and summarized by human drivers and utilized to evaluate the AI's capabilities under these scenarios.

Mapping Capabilities. We use a list of tuples (m_i, s_i, a_i) to represent the capability regarding task **scenarios**, which is both explainable for human and readable for AI. For each tuple, m_i represents the team member (human or AI), s_i denotes the type of task in a specific scenario, and a_i denotes the corresponding accuracy of completing the task in the scenario s_i for m_i. The **scenario** s_i is

represented by a set of task-related **conditions** c_i summarized by human or AI that greatly affect the accuracy and exhibit significant differences in capabilities of completing tasks. Specifically, s_i could be binary vectors, i.e., $s_i = [c_1, c_2, ..., c_n]$, where $c_1, c_2, ..., c_n$ is 0 or 1 to denote whether current condition exists in scenario s_i. And the n varies for different human-AI cooperation tasks, depending on the possible number of conditions that human and AI could summarize that affect task accuracy. Here, it is notable that the notion of **condition** also shares the similar connotation of "features" in feature visualization or feature importance of explainable machine learning algorithms [28, 87] as we mentioned above, which, from AI's perspective, could be treated as AI's summary of conditions that could make sense to human observers. For example, in driving tasks, assume there are three conditions summarized by human and AI in total: "dark-night vision", "vision is partially obscured by large obstacles" and "vision is blurred". And, a dark-night vision blurred by rain may make it much harder for AI to accurately recognize the surroundings and thus decreases the accuracy of making correct actions. Hence, the scenario here is "dark-night vision" and "vision is blurred", which could be represented as $s_1 = [1, 0, 1]$, and the tuples corresponding to the scenario are (*AI*, [1,0,1], *low*) and (*human*, [1,0,1], *high*). As shown in Fig.3, these tuples are collected from both the human side to form H (human mental model) and the AI side to form A (AI mental model). Human can leverage their expert knowledge to summarize the scenarios in which their prediction is influenced while AI can utilize the accuracy statistics of previous tasks to break down the cases by exploring the combination of a set of conditions. These representations are gradually updated and revised along with the cooperation to keep a more accurate mapping between members' capability and corresponding task scenarios as the mutual understanding goes deeper [83]. Based on these representations, the original tasks can be further grouped into fine-grained task scenarios for assignment.

Human-centric capability representation. For most human-AI cooperation, more attention should be paid when AI tends to fail and then hands over to the human. Hence, we focus on tuples that highlight scenarios where human has higher accuracy than AI, i.e., tuples where m_i is human, a_i is relatively higher than AI. Therefore, we could utilize the scenarios s_i of these tuples to form a human-centric capability representation, denoting task groups suitable for handing over to the human. In this way, the H and A could be simplified as list of s_i^H and s_i^A. Each list includes selected scenarios from AI or human where AI performs worse and requires handing over to human. Such as in Fig.3, the evaluation from human as (*Human*, [1,0,1,0], *High*) is recorded and simplified as [1,0,1,0] in s_i^H. The advantage of doing this lies in two folds: it simplifies the redundant representations of the mental models, as well as relieves human's cognitive workload in cooperation.

As illustrated in Fig.3, the lists of tuples (or binary vectors of scenarios) output by human and AI after task grouping can be regarded as their initial and individual mental models for TAHAC. And negotiations will be placed to achieve a consensus about the shared grouping between both sides and assign the task considering each member's strengths and weaknesses according to various task scenarios.

3.3 Negotiation

Negotiation is the process to resolve the differences between H and A by merging them through negotiation methods g and reach a shared mental model SMM, i.e, $SMM = g(H, A)$, that enables further task assignment. Usually, negotiations are required to catch up with the latest mental models as the members gain a deeper understanding of each other. To merge human mental model H and AI mental model A in forms of tuples (m_i, s_i, a_i), the considerations should include aligning each m_i, s_i, and a_i in H and A to form agreed tuples to be mapped into task assignment. However, for TAHAC, as discussed in task grouping, we focus on scenarios which should be handed over to human and utilized human-centric capability representation to highlight these scenarios. Hence, negotiation could further be simplified to only align scenarios where human or AI believe human's capability is superior to AI. That is, the negotiation methods g should realize $s_i^T = g(s_i^H, s_i^A)$, as shown in Fig.3. Since the scenarios are actually binary vectors which represent the presence or not of certain conditions, we designed to utilize the bitwise operations as feasible negotiation methods. More specifically, we propose to utilize two basic strategies, i.e., bitwise AND (&) and bitwise OR (|) to merge two different scenarios into a shared one. Noting that usually the human side should be the actual controller of the whole system and their grouping results have higher priority [72], the proposed merging strategies during the negotiation are further designed to lean towards human.

Human-biased Bitwise AND. Bitwise AND is the logical operation of computing AND for each bit of the binary vector s_i proposed by the human and AI. In detail, for s_i^H (e.g., $[1, 1, 0]$) from human side and s_i^A (e.g., $[1, 0, 1]$) from AI side, the shared scenario is computed as $s_i^T = s_i^H \,\&\, s_i^A$ ($[1, 0, 0] = [1, 1, 0] \,\&\, [1, 0, 1]$). If after bitwise AND operation, s_i^T is a null vector in which all the elements are zero, we will let $s_i^T = s_i^H$ instead, based on the assumption that human's assessment is more privileged. This method is called human-biased bitwise AND during evaluation. Since in this way, fewer conditions are included in one scenario after negotiation (e.g., $[1, 0, 1]$ specifies tasks satisfying two conditions, while $[1, 0, 0]$ specifies tasks satisfying one condition), it makes a mild strategy that generates a more relaxed scenario covering a broader range. More cases will hit the scenarios and be assigned to human according to the shared mental model.

Human-biased Bitwise OR. Bitwise OR is the logical operation of computing OR for each bit of the binary vector s_i proposed by the human and AI. It is computed as $s_i^T = s_i^H \mid s_i^A$ ($[1, 1, 1] = [1, 1, 0] \mid [1, 0, 1]$). Similarly, if s_i^T equals to neither s_i^A nor s_i^H, s_i^T will be replaced with s_i^H to lean toward human's assessment. This method is called human-biased bitwise OR during evaluation. In this way, more conditions are included in one scenario after negotiation, it forms a strict strategy that generates a more rigorous scenario and drives the system to be careful and safe, i.e., fewer cases would satisfy the scenarios the shared mental model suggested but each of them is paid more attention to.

After the negotiation processes, the merged representation SMM can then be utilized for subsequent task assignment. Based on our human-centric representations, subsequent tasks that satisfy the scenarios in s_i^T are more suitable to be handed over to human for

scrutiny and completion. If human or AI finds the current *SMM*-based task assignment inappropriate during cooperation, such negotiation processes can be carried out iteratively to continuously update *SMM* based on the updated *H* and *A* of the human and AI, thus forming the revised task assignment strategies.

3.4 Evaluation Metrics

To assess the effectiveness of human-AI cooperation and explore the cooperation dynamics with the introduction of CASMM, we introduced the **Cooperation Score** and **Assignment Strategy Accuracy** as additional evaluation metrics aside from accuracy and time efficiency.

Cooperation Score: We design the cooperation score to measure how well the shared mental model mechanism facilitates the task assignment in human-AI cooperation. The metric is computed in the task loop to serve as an indicator for fixing the potential issues in the cooperation. Different from an ordinary task completed by a single human or AI, in the cooperation scenario, accuracy is not sufficient to evaluate the quality of the shared mental model. Consider two extreme cases: in the first case, the human member is sophisticated and does not care about the suggested assignment at all. Therefore, the human completes all the tasks alone with high accuracy regardless of the task assignment. In the second case, the human member obeys the task assignment, but unfortunately, the assignment strategy given by the shared mental model is terrible and the human is always given tasks she is not good at. Both cases are bad cooperation scenarios since they waste the resources of the member's strength. Therefore, to evaluate how well the shared mental model mechanism performs, three key factors should be taken into account, i.e., assignment acceptance, assignment strategy, and the accuracy of the task.

For an ideal cooperation practice, the human should be assigned with tasks that she is good at completing (correct assignment strategy), and she should agree with the assignment and be willing to complete her share given by the shared mental model (high assignment acceptance). Finally, the task should be correctly completed (high accuracy). Any factor missing leads to a decrease in cooperation quality, while the correctness of completing the tasks weighs more. Hence, the cooperation score is computed as follows:

$$Cooperation\ Score\ = s \times 0.5 + a \times 0.5 + c \tag{1}$$

where, for each cooperation task, s denotes the correctness of the current assignment (+1 for correct; -1 for incorrect); a is the assignment acceptance of the current task (+1 for accepted; -1 for not accepted); c represents the correctness of the result (+1 for task being completed correctly; -1 for incorrectly). The correctness of the current assignment could be computed as: if the assigned member finishes the task correctly, the assignment is correct. Otherwise, it is incorrect. The score ranges from -2 to 2, where a higher score indicates more successful cooperation. An ideal cooperation process earns 2 points when the task is correctly assigned, the assignment is accepted, and the result is correct. If the result and assignment are correct, but the human rejects the assignment, only 1 point is obtained to punish the waste of team resources. If both the cooperation acceptance and the strategy are incorrect, the team's score is at most 0 since the cooperation fails in this case. The cooperation score is computed for each task and the averaged cooperation score can empirically reflect the quality of the shared mental model.

Assignment Strategy Accuracy: We also introduce the assignment strategy accuracy to theoretically measure the accuracy of the task assignment by computing its similarity with an ideal assignment strategy. The ideal assignment can be obtained by exploring the performance statistics with ground truth and summarizing scenarios that are worse for AI to handle than the human user it's teaming up with. Given the scenarios being presented using binary vectors, we can utilize the Jaccard similarity (which is widely utilized to compare the similarity of binary arrays) to compute the similarity between the current assignment strategy and the ideal strategy as assignment strategy accuracy. An accuracy closer to 1 means a higher similarity to the ideal one and makes a better assignment.

4 CASMM: A PROOF-OF-CONCEPT APPLICATION

4.1 Research Questions

To evaluate the feasibility of the proposed capability-aware shared mental model, the current research conducted a study that examined the following research questions:

RQ1: How would capability-aware shared mental model for task assignment affect the efficiency of Human-AI cooperation compared with no task assignment or human-generated task assignment?

We conduct an experiment of 3 phases(exp 1-3), comparing task assignment based on the capability-aware shared mental model, no task assignment, and human-generated task assignment. To quantify cooperation efficiency, we measure the success rate and task completion time for each experiment group.

RQ2: How would the task assignment plan generated from exp 1-3 be different from the optimal task assignment plan?

We recorded the capability of each team member for each experiment and calculate the optimal task assignment plan according to their actual corresponding accuracy for all possible classification of tasks as ground truth and then compare it with other exp 1-3 task assignment plans. We hypothesize that the CASMM is closer to the ground truth.

RQ3: How would the negotiation methods influence the shared mental model come up by human-AI teams?

We compare the 2 types of negotiation methods (human-biased bitwise AND (&) and bitwise OR (|)) regarding the process of building CASMM and deriving Task assignment plans compared to the ground truth.

RQ4: How would building the capability-aware shared mental model help humans understand AI's actual capability for a specific task and then become more confident about the task assignment plan?

For the experiment sessions, we also asked the participants to finish a post-experiment survey to collect their feedback specific on their perception of the CASMM based task assignment plan, their experiences with human-AI cooperation, etc.

4.2 Task Design

First, we conducted a lab study task for our user test that focused on problems of task assignment for human-AI cooperation. We

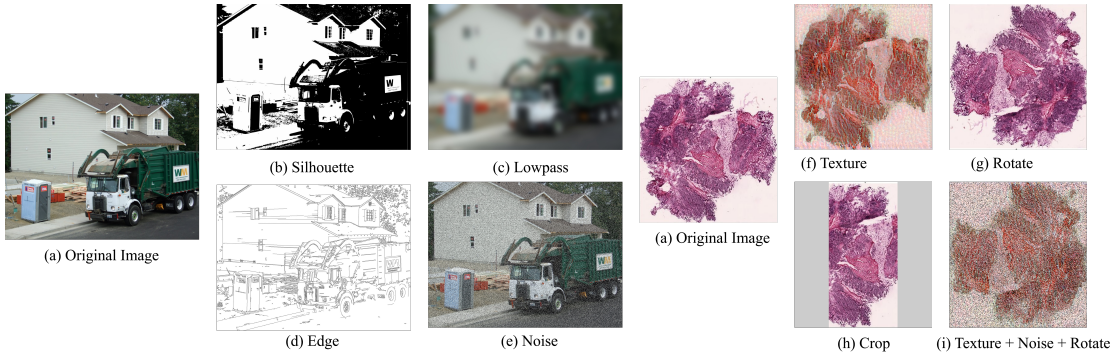

Figure 4: (a) Example pictures from automated driving and medical diagnosis tasks. They are the original picture to perform the post-processes we chose during the experiment. The post-processed pictures with our chosen techniques to simulate real-world scenarios. (b) Silhouette. (c) Lowpass. (d) Edge. (e) Noise. (f) Texture. (g) Rotate. (h) Crop. (i) Texture + Noise + Rotate (a multi-condition example).

designed an image labelling mission as the task for building a capability-aware shared mental model during human-AI cooperation. This mission takes on a similar form to lots of human-AI cooperation tasks we mentioned that involve image recognition, such as medical diagnosis and automated driving, but simplified for testing and evaluating. In the experiments, we present pictures for human-AI teams to categorize into one of the 20 categories. Along the experiment, human and AI need to come up with a task assignment plan that determines which kind of images is better to be assigned to which member to label, in order to achieve high labeling accuracy and time efficiency. We selected images from 12 out of 20 categories, with the rest serving as confusing options for both human and AI. These 12 categories include barrel, bottle, bird, dog, box, lion, pan, rocket, sailboat, snake, van, and printer from ImageNet [65], covering animals, furniture, transportation, and appliances to reach a wider range of real-world objects [21]. The original pictures are chosen by two authors to ensure that they are clear and relatively recognizable. The 8 confusing options are derived from the categories whose appearances are similar to the selected images, such as including the "cat" label as a confusing option corresponds to the selected "dog" or "lion" label. All of the original images are then post-processed.

Gaining inspirations from real-world tasks such as *automated driving* and *medical diagnosis*, we selected 7 representative image post-process techniques. Through these techniques, we produced pictures with specific characteristics proven to affect both AI and human's performance in the above two real-world missions. These characteristics may be further explored, clarified, and utilized between human and AI to distinguish the tasks, thus forming a capability-aware shared mental model. The post-process techniques are applied to the pictures with the same parameters that ensure the stability of the difficulty for most human workers and AI. Furthermore, as in real-world scenarios, these basic post-process techniques could be combined, producing pictures with the compound characteristics (these combinations are carefully selected for clear distinction, such as shown in Fig. 4 (i)). The post-processing techniques and their corresponding reasons for being selected are described as follows, corresponding to Fig.4:

(1) Silhouette: the pictures are posterized using Portrace [70] to obtain the silhouette version of the original pictures. It simulates the malfunction in the automated driving scenario when the shutter does not allow the entrance of the correct amount of light through the lens [69], or when the environment is dark outside, as shown in Fig. 4 (b). It might cause image detection algorithms for automated driving to fail, as well as influence the human drivers' judgement.
(2) Lowpass: the blurred version of the original picture is obtained using the function "Gaussian filter" embedded in PIL package in python according to the size of the picture. It simulates the blurred vision during automated driving accrued when the camera is out of focus, or in environments such as rain or fog, as shown in Fig. 4 (c). This can likewise affect effective driver judgment and the outcome of autonomous driving image recognition.
(3) Edge: the pictures' edges are extracted using the Canny algorithm embedded in OpenCV package. It also simulates the malfunction in the automated driving scenario when the shutter does not allow the entrance of the correct amount of light through the lens [69], or when the environment is suddenly turning bright outside, as shown in Fig. 4 (d).
(4) Noise: random salt and pepper noise is applied to the pictures, as shown in Fig. 4 (e). It simulates the error during the noise reduction phases, which makes RGB cameras for automated driving cannot remove the noise correctly [69]. The excessive amount of noise would interfere with the algorithm and the driver's judgment.
(5) Texture: the pictures are stylized by transferring another texture picture's style into the original picture [19]. It simulates the scenario in medical diagnosis that different stains and their concentration's effect on the judgment of the physicians and the automated diagnostic algorithms in the diagnosis of pathology images [39], as shown in Fig. 4 (f). Generally, such images require image normalization for further diagnostic processing.
(6) Rotate: the original pictures are rotated by 180 degrees. It simulates the rotation of the tissue to be observed due to

production errors in pathology image diagnosis, as shown in Fig. 4 (g). This can affect the physician's ability to find the tissue to be observed, as well as the accuracy of some image recognition networks that perform medical diagnostics [80].

(7) Crop: the middle parts of the pictures are cropped and the rest area is filled with colors that are not consistent with the cropped edge (preferably black or white) for more distinction, as shown in Fig. 4 (h). It simulates the inability to observe intact tissue in pathology images due to improper manipulation resulting in parts of the tissue being folded, obscured, or stained by ink, etc. This again affects the diagnostic quality of the physician as well as the machine learning algorithm [39].

Therefore, for simplification, we only need to record opinions where scenario s_i in tuple (m_i, s_i, a_i) is expressed using the combination of the above basic 7 conditions. This simplification is acceptable since it requires a low threshold for human to distinguish these conditions with little dissent, hence leading to fast convergence. And, AI is able to recognize these scenarios too, because recognizing the type of these process techniques is a standard task commonly used in the field of computer vision as the auxiliary information [9, 75]. Furthermore, the selection of these 7 conditions is corresponding to [20]'s findings that current image-detecting neural networks are biased towards these similar conditions like "texture" and "edge" in accuracy. In this way, if the AI has an evaluation as "I have an accuracy of 88.2% while labeling pictures processed with a combination of edge and rotate", we only need to record "$(\{edge + rotate\}, AI, 88.2\%)$" in A to represent such evaluation. As we mentioned before, for the experiment, we simplified such evaluations to focus on scenarios when tasks should be handed over to the human. Hence, one step further, these opinions could be recorded as a series of binary vectors of length 7. For example, if the scenarios are organized in the order as in Fig.4, $[0, 0, 0, 1, 0, 0, 1]$ could represent that pictures post-processed with edge + rotate techniques should be handed over for human to label instead of AI. In the real experiment, we applied six combinations of post-process techniques to these pictures, which include lowpass, stylized + edge, silhouette + noise, crop + rotate, silhouette + stylized + rotate, rotate + edge + noise.

For comparison, there are 2 kinds of AI helpers that a user might encounter as a team member during the user study, they share the same capability (or accuracy) of labeling these images and identifying the combination of conditions of an image, but adapting different techniques while forming A. We trimmed a pre-trained convolutional neural network based on VGG19 [73], to reach distinctive accuracy for the six process types. For each process type, fixed recognition accuracy is given to the AI to simulate its diverse capabilities of identifying images with varied conditions. However, the AI's ability to express A varies, one kind of AI could only compute and give opinions using one of the seven conditions. The other kind of AI could actively combine conditions to compute and compare the accuracy of different post-process while forming A. For example, the second kind of AI could record "$[1, 0, 0, 0, 1, 0, 0]$ (texture + silhouette)" as a set of s_i, while the first kind of AI could only record "$[1, 0, 0, 0, 0, 0, 0]$ (texture)" or "$[0, 0, 0, 0, 1, 0, 0]$ (silhouette)" according to its comparison. Apparently, the second kind of AI has a better express ability of A than the first kind.

After careful selection, post-processing and training process, we obtain our pictures and AI agents for the human-AI cooperation task assignment. We conducted a pilot study that randomize the order of the pictures we chose for labeling for several users to label, and ensured that the pictures post-processed with the same techniques are relatively similar in difficulty. To realize the process of the designed workflow shown in Fig.2 on our image labeling task and conduct contrast experiments, we designed a prototype system and user study as follows.

4.3 User Study

We designed a three-phase user study in our prototype system, which will be called experiment 1-3 (exp 1-3). We recruited 40 volunteers for our exp 1-3. The participants are students from a local university, aged from 21 to 29, and have a relatively equal gender proportion of 18 women and 22 men. And according to the pre-survey, they all have some sense of how AI works in categorizing pictures but are not informed of AI's behavioral patterns or algorithms in our experiment. During three formal experiments, users will be asked to label 120, 180, and 180 images in a fixed order, with each cooperation session (as in 2) containing 30 images. Pictures processed with the same techniques appear relatively in the same proportion for each cooperation session of the experiment, to ensure human and AI both encounter enough cases of each process and discover them. Also, we conducted a pilot study to ensure that for our experiment 1-3 at this length and difficulty, the users' learning effect can be ignored.

The three formal experiments are completed in chronological order with a time lag of at least 1 day by every user to avoid influence among experiments. Users are free to choose the time to complete three experiments in order to avoid the influence of the previous experiment. All of the users are familiar with the interfaces, interactions we designed, and the several basic picture conditions we selected in advance, to avoid the hampering of the unfamiliarity of the system during experiments. And following the completion of all 3 experiments, each user was asked several questions about the experience difference, the system preference and whether they would like to work with AI this way in the future, etc. The participants were paid twice the average hourly earnings of local areas per hour after the experiment, according to their actual participating hours.

In the first experiment, the users are required to label 120 pictures and are able to acquire the correct labels after the submission of their choices. And after completion, they are asked to self-evaluate the process types of the pictures that are hard for them to recognize, using the combination of our proposed seven conditions. The AI also finishes labelling these pictures in the background and evaluating itself, thus forming its own self-evaluation, also using seven proposed conditions.

For the second experiment, the AI and user are required to label another 180 pictures together with only the human's mental model H (HMM) for task assignment. Human-AI teams will enter negotiation sessions after every 30 pictures of cooperation sessions, but here we only record human's opinions, which represent H for the task assignment of the next cooperation session. The cooperative

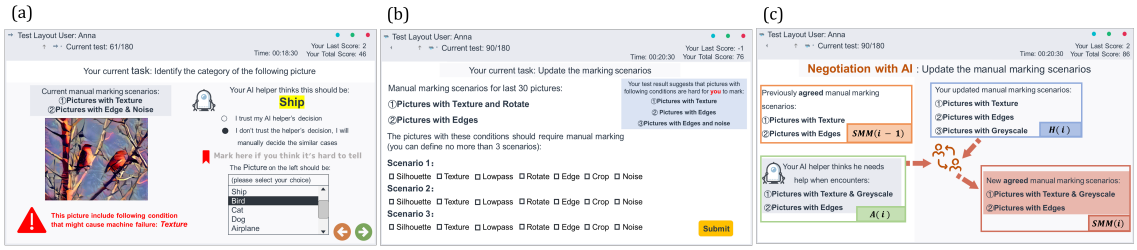

Figure 5: (a) The interface design of cooperation sessions, which incorporates H or SMM above the images and utilizes the warning sign to represent a mild form of work assignment. Current H(or SMM) include two agreed scenarios: "texture", and "edges + noise", and the current image is post-processed with "texture", hence it is highlighted with a warning sign and an explanation illustrating that it's highlighted because it's processed with "texture". Here the scores are what we described as cooperation scores. It is shown for the sake of recording. Since the meaning of cooperation scores is not provided to users, they are only requested to cooperate with AI as we instructed and correctly label the images without referring to the scores. (b) The interface design of negotiation sessions to express, and store human's mental model on current team capability. (c) The interface design of the pop-up window representing negotiation sessions, showing the process of forming $SMM(i)$. Here we highlight and mark the corresponding elements of building shared mental model in the paper (negotiation, $SMM(i-1), A(i), H(i)$, and $SMM(i)$). The highlights and marks are invisible to users in the experiment.

labeling of the pictures follows the interface as shown in Fig.5(a). In these interfaces, human-AI teams would not refer to the correct answers of the current images, simulating the real-world occasions that human-AI cooperate on tasks without a correct ground truth to refer to. And the H of the current session (annotated by the human during the last session) is shown above the images. And based on H, we highlight the images with a warning sign when the current task satisfies the H. This counts as a mild form of the work assignment: if an image in our interface is placed a warning sign, according to current H, the image is more suitable for distributing to human to label, according to human's judgement. The warning sign would suggest human to choose "I don't trust AI" under these cases. However, for the last cooperation session of 30 pictures (from the 151^{st} to 180^{th} pictures), we enter the "Quick Sort Session". It's a more radical way to test our human-AI work assignment method: only the pictures satisfying H chosen by the end of the 120^{th} picture are assigned to human workers, with AI labelling the rest of the images as the answer of the human-AI team. "Quick Sort Sessions" adapt relatively extreme ways of work assignment. We adapted this rather "radical" assignment strategy to explore the possibilities, and also temped to amplify the boost of a capability-aware shared mental model than the human mental model. It could also serve as a reference for the worst cases where consulting a human expert is very expensive.

For the third experiment, we generally follow the same process that experiment 2 has, but utilize the shared mental model for task assignment. The AI and user are still required to label another 180 pictures, enter negotiation sessions after every cooperation session of 30 pictures, and complete the final session of the experiment following the "Quick Sort" mode. However, in experiment 3, we considered both H and A to form shared mental model (SMM) among team members, thus containing the pop-up window designed in Fig.5(c) for every negotiation session. Here, human and AI work together to update the SMM. The system uses our proposed negotiation methods to update SMM in our interface design shown in Fig.5(c). During experiment 3, we compared human-biased bitwise AND (&) and bitwise OR as negotiation methods to form SMM, as mentioned in section 3.3. Also, we compare the influence of the two kinds of AI's ability to express A, as described in section 4.1. Hence, we randomly divide users into groups of 10, with each group adapting one of the two methods and collaborating with one of the two kinds of AI.

After the three experiments, we conducted short interviews with the participants, including three scale questions and a short open dialogue, to understand their perceptions and experiences of working with the AI to form a shared mental model during the experiment. The three scale questions asked the user to rate experiment 2 and experiment 3 on a scale of 1 for strongly disagree, 7 for strongly agree, and 4 for neutral. The three statements included: I am confident in the task assignment formed during the experiment; I have a better understanding of different situations in which the AI and I are prone to make mistakes; and I have a better understanding of the capabilities and limitations of the AI algorithm.

4.4 Results

We collected accuracy, session time length and cooperation scores along with the detailed user log throughout the experimental process of the three experiments. For comparison, we divided our user-AI teams' performance into the following seven cooperation modes: human-alone mode (marked as "Human"), AI-alone mode (marked as "AI"), human + AI mode (human makes the decision of labeling while AI's labels are always available to humans, without any task assignment, as in first sessions of exp 2 and exp 3, marked as "Human+AI"), human + AI + human mental model mode (during experiment 2, marked as "Human+AI+HMM"), human + AI + capability-aware shared mental model mode (during experiment 3, marked as "Human+AI+SMM"), quick sort + human mental model mode (last session of exp 2, marked as "Quick Sort+HMM") and quick sort + shared mental model mode (last session of exp 3, marked as "Quick Sort+SMM").

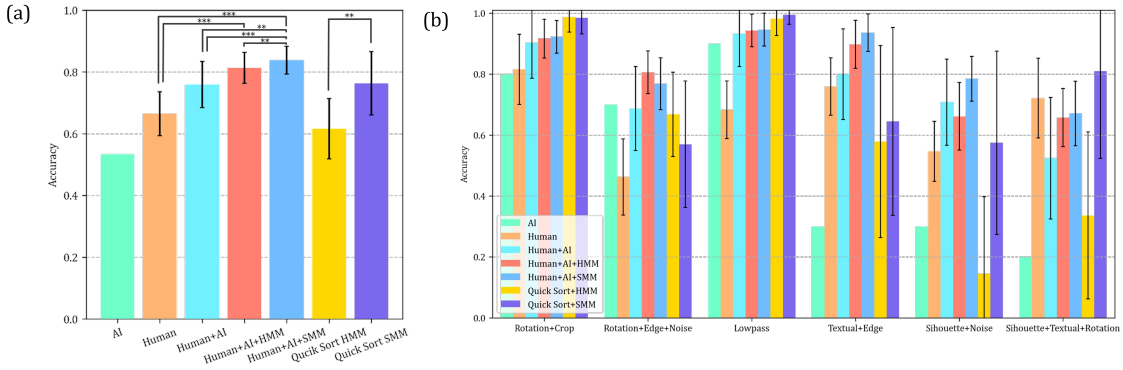

Figure 6: The plots of comparing the accuracy of seven different work modes from different perspectives. (a) Overall accuracy of 7 involved work modes. The colored bars indicate the mean accuracy of the corresponding work mode, and black lines in the middle extend to the upper quartile and the lower quartile of the data (= p < .05, *** = p < .001, Mann-Whitney test). (b) The breakdown of average accuracy into scenarios (the combination of post-processing conditions) of 7 work modes.**

4.3.1 RQ1: How would CASMM for task assignment affect the efficiency of Human-AI cooperation compared with no task assignment or human-generated task assignment?

Our first research question addresses the effectiveness of our proposed capability-aware shared mental model to facilitate TAHAC tasks. We computed the average accuracy of human-AI teams correctly labelling the pictures under the above 7 working modes in Fig.6(a) and compare the average time per task under six work modes (since the average time of AI completing the tasks is negligible) in Fig.7(a).

Accuracy-wise, as in Fig.6(a), it is worth noting that work modes with a mental model (either HMM or SMM) gain a rise in accuracy compared with the human+AI mode where AI decisions are only references to human. Although the accuracy of AI stays the same across work modes, introducing the mental model into cooperation and utilizing a warning reminder to embody the mental model do improve the team performance in terms of accuracy. The work modes that human-AI teams work with shared mental model outperform the modes with pure human's mental model ($p < 0.05$), and we believe it is due to the negotiation between human and AI bridging to a more accurate shared mental model than human mental model. Even in the "Quick Sort" sessions where human are only given images selected by the shared mental model, it achieves higher accuracy than human+AI ($p < 0.001$).

On the other side, time-wise, as in Fig.7(a), human+AI teams with mental models require much less time per task than those modes without mental model, with SMM reducing over 30% of the average time compared with human-alone mode ($p < 0.001$) and nearly 25% with the human+AI mode ($p < 0.05$). Moreover, for "Quick Sort" modes, since the pictures not qualifying the HMM/SMM are skipped for human and completed by AI in the background, the average time is largely reduced and even probably can be zero if the human completely trusts AI in labelling all the pictures. Given the average accuracy of the quick sort modes are competitive compared with the human-alone mode (HMM) and even the human+AI mode (SMM), it could be used as a trade-off to balance the accuracy and the time, taking much shorter time to achieve relatively high accuracy.

We also compare the breakdown of accuracy into all the occurring scenarios during experiments under different work modes in Fig.6(b). For most scenarios, the human-AI teams with shared mental model outperform the other modes since it enables human and AI to negotiate the opinions of confidence based on shared scenarios. For pictures processed with "Silhouette + Texture + Rotation", there is a significant accuracy decrease in human+AI mode compared with the human-alone working mode. This may be due to the fact that the extremely low accuracy of AI brings plenty of distractions for humans, resulting in more uncertainty in decision-making. This is also close to the results mentioned in Bahrami et al. [2], since human and AI's visual sensitivities in such scenario have a larger difference, resulting in "two heads worse than the better one". Yet we can see that modes with human mental model and shared mental model help fill in such decrease during the cooperation process, making the accuracy closer to the high accuracy of human-worker alone. And for those categories where AI performs worse than human workers in particular, work modes with shared mental model outperform more than the modes with human mental model. The above observation indicates that utilizing the shared mental model is of great help for cooperation not only when the AI is quite accurate but when the AI is extremely weak as well.

Hence, in general, building SMM for TAHAC does help improve the efficiency of human-AI cooperation, compared with no task assignment or the task assignment accord with the human's mental model. It shows a promising time-accuracy balance to boost human-AI cooperation.

4.3.2 RQ2: How would the task assignment plan generated from exp 1-3 be different from the optimal task assignment plan?

Our second research question aims at evaluating the validity of the task assignment plan that human and AI build together through the shared mental model we proposed. Hence, we utilized the proposed assignment strategy accuracy described in section 3.4 for such evaluation. It is computed by comparing the statistical "best task assignment strategy" for each human-AI team with their actual task assigning plans formed during experiments.

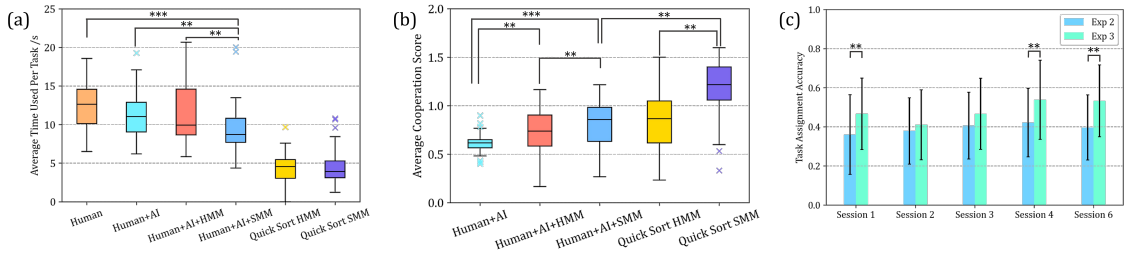

Figure 7: (a) Comparison of average time per task under 6 involved work modes. (b) Comparison of average cooperation score per image under 5 involved work modes. (c) The assignment strategy accuracy of five negotiation sessions that human-AI teams entered during experiments along the session (during session 5 the task assignment plan is not updated for comparing quick-sort modes) (= p < .05, *** = p < .001, Mann-Whitney test).**

As in Fig.7(c), we could observe an overall rising trend for the assignment strategy accuracy as more negotiation sessions entered along the process of our experiment. And the task assignment based on the shared mental model (experiment 3) gained higher assignment strategy accuracy than that based on the human mental model (experiment 2). The process of forming a capability-aware shared mental model through negotiations helps the human-AI teams to discover their team capabilities more comprehensively and thoroughly. Thus, it's beneficial for leading toward more rational work assignments with higher assignment strategy accuracy. This also indicates that the iterative process we implemented is able to help build a shared mental model that leads towards an assignment of work close to human and AI's real capabilities, with relatively few iterations. Interestingly, there exists a mild decrease in assignment strategy accuracy during the last session (by the end of the experiment) for both experiment 2 and 3, it is probably because human-AI teams focus on revising current mental models rather than introducing new perspectives of capability evaluation as more sessions countered.

4.3.3 RQ3: How would the negotiation methods influence the shared mental model come up by human-AI teams?

This research question explores the possible influences of the negotiation methods we designed on the human-AI shared mental model. Therefore, we compare the assignment strategy accuracy of four groups of human-AI teams we divided during experiment 3, which adopt different negotiation methods or interact with AI with different express ability of its opinions (A). As in Fig.8, the average of human-AI teams' task assignment scores are generally higher when machines have the better express ability of A (Multi-conditions). And when the negotiation sessions adapt human-biased bitwise AND method, the distribution of task assignment scores is more concentrated and shows a clearer rising tendency along the workflow session-wise. In general, using the human-biased bitwise AND method with an AI with a better ability of expressing A to negotiate helps human-AI teams come out with more accurate mental models with concentrated distribution.

4.3.4 RQ4: How would building the capability-aware shared mental model help humans understand AI's actual capability for a specific task and then become more confident about the task assignment plan?

Our fourth research question aims at considering whether human becomes more aware of the AI's actual capability, thus more actively cooperating with AI through building the shared mental model. Hence, we compare two aspects of the results: the cooperation score and subjective feedback.

The cooperation scores per task under 6 work modes are shown in Fig.7(b) (Since human/AI-alone work modes do not involve cooperation, they do not produce cooperation scores). The cooperation score is introduced previously to additionally evaluate the effectiveness and efficiency of the human-AI cooperation. Clearly, human-AI teams with mental models take on higher cooperation scores than those without a mental model ($p < 0.001$). The result demonstrates the mental model helps push better human-AI cooperation tendency, and assist human to interact and cooperate with AI more properly. SMM shows higher cooperation scores than the HMM, indicating that forming a shared mental model through negotiation sessions makes a more efficient advancement for human to understand more about AI's real capability. Note that in the quick sort mode, the shared mental model largely surpasses the others in scores, which reveals that the active involvement in negotiation helps participants get high cooperation with accurate strategies.

Subjective scale questions' distribution is shown in Fig.9. Generally, both experiment 2 (M = 4.60 on a 7-point scale) and 3 (M = 4.83) help participants to realize the capability of AI agents. However, participants were more aware of the ability difference between the AI algorithms and themselves during experiment 3 than experiment 2 ($p = 0.001$). Also, participants report they are more confident of the negotiated shared mental model (M = 4.97) in experiment 3 to make correct assignments than their own mental model (M = 4.33) in experiment 2 ($p = 0.023$).

Similar opinions are expressed during open dialogues. Around 2/3 of the total users (29 out of 40 users) prefer the cooperation mode in experiment 3 working with AI as a team. And more than 70 percent of the users expect such type of cooperation with AI on real tasks in the future. For instance, "I like making decisions together with an AI helper. It shows a greater sense of interaction (P19, Female)"; "There are lots of post-processing techniques involved and AI agents might fail in multiple occasions. But with the combination of AI's scenarios with mine, I could find occasions that AI usually make mistakes more easily (P35, Male)".

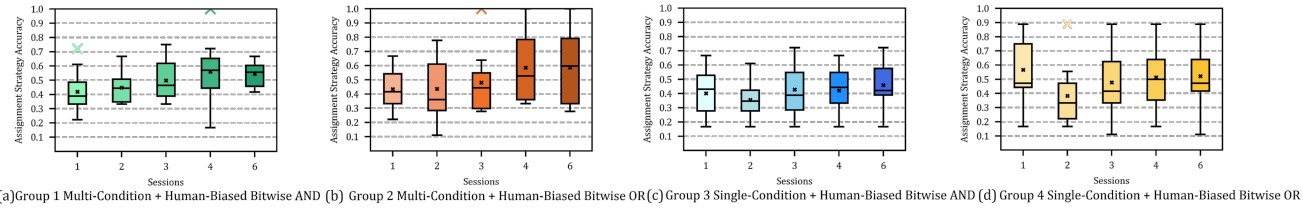

Figure 8: Comparison of the assignment strategy accuracy of four groups of human-AI teams across different sessions (30 pictures per session) in experiment 3. The "Single-condition" and "Multi-condition" represent the AI's capability of expressing A, which corresponds with the first kind of AI and the second kind of AI in section 4.1. The negotiation method is accord with section 3.3. The mark "x" in the box denotes the mean value and the solid line in the box denotes the median value. (a) The similarity of the user group with A using multiple combinations of conditions and human-biases bitwise AND method. (b) Similarity of the user group with the same A as group 1 and human-biased bitwise OR method. (c) Similarity of the user group with A using single conditions and human-biases bitwise AND method. (d) Similarity of the user group with the same A as group 3 and human-biased bitwise OR method. Each subplot contains data from 10 users of the group.

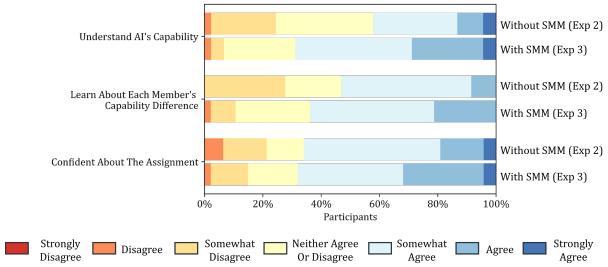

Figure 9: The scale distribution of subjective questions. The three statements on the left side of the figure correspond to each of the three scale statements we proposed. Understand AI's capability: I have a better understanding of the capabilities and limitations of the AI algorithm; Learn about each member's capability difference: I have a better understanding of the situations in which the AI and I are prone to make mistakes; Confident about the assignment: I am confident in the task assignment formed during the experiment.

5 IMPLICATIONS, LIMITATION AND FUTURE WORK

5.1 Design Implications

Improving the trust between human-AI cooperation: interestingly, we notice that involving negotiation in the workflow of forming the shared mental model mitigates the trust issues to some extent. 70% of the participants (28 out of 40) agree that they feel more trusting in AI during exp 3 than exp 2, where participants build and revise shared mental model with AI through the negotiation process. Participants are observed to trust the shared mental model they built and obtained a clear understanding of the AI's decision. From the quantitative perspective, the cooperation score of the work mode with a shared mental model is much higher than the mode with human mental model and human only refer to AI mode. Since the cooperation score is directly influenced by the acceptance of the grouping, which reflects the trust of the current task assignment, it can be inferred that more trust spreads between participants and AI as the shared mental model forms. On the other hand, based on the subjective feedback, it is worth noting that most of the participants report to be more aware of when AI is likely to make a mistake, thus they become more confident of the cooperation results. This is consistent with our observations during the experiment: participants take less time questioning the AI's answers, but tend to complete tasks relying on the warning signs according to the shared mental model.

The applicability of capability-aware shared mental model: As described in Sec.4.1, our task design is based on several characteristics of the real-world human-AI cooperation tasks. We simulate and simplify the work assignment problems that human-AI teams in practical tasks might encounter. Hence, similar tasks we mentioned like medical diagnoses or semi-autonomous driving can directly adapt our system for building and utilizing shared mental models with a few modifications. Autonomous driving involves dynamic image recognition during actual operation, then the content being intercepted for analysis can be a short video or a key-frame where the current road conditions change. If the current system is used for medical diagnosis, sometimes the doctors and AI have to perform the task of locating the lesion. Hence, when defining medical image conditions for shared mental model, the human doctor or AI can provide the corresponding medical knowledge basis relating to making the diagnosis, and extract some common conditions of this knowledge to obtain a more accurate task classification to guide the task assignment. For these tasks, s_i in our tuples to describe shared mental model could utilize conditions like "length of largest tumor ≤ 0.13" [83] or "rainy foggy environment + obstructed view" as scenarios. Also, such a shared mental model is not only limited to human-AI cooperation for image-based tasks. Using other dimensions of task features (e.g., numeric features, text features, etc.) for task classification to form shared mental models can also improve the efficiency of human-AI cooperation. For example, for the review of loan approval, some existing challenges include the difficulty for human operators to determine whether all the automatically approved lenders meet specific conditions for loan origination due to distrust and lack of understanding of the automated approval process or algorithms. We could utilize certain dimensions of lender characteristics as conditions to form

our proposed capability-aware shared mental model during the cooperation process. In this way, the human operators can quickly understand the AI algorithm's capabilities and form a task assignment based on human-AI's consensus, identifying risky lenders that the AI algorithm might overlook during the lending process. For example, a lender with fair assets but a potentially risky investment may not be able to repay the loan, but AI might approve such lenders due to their fair assets. Hence, the combination of "fair assets + risky investments (which can be described by specific data or corporation, etc.)" can be utilized to form the capability-aware shared mental model.

Currently, the definition of "condition" and "scenario" seems human-centered, which has been demonstrated to be effective. Allowing the AI to define the scenarios based on its decision boundaries may also be a feasible approach. Works in the area of feature visualization and AI behavioral interpretability and explainability will further facilitate AI's proactiveness in proposing some conditions or scenarios that make sense to human from their perspective [5, 6]. Such efforts may further facilitate and promote cooperation between human and AI with CASMM, or contribute to future various forms of human-AI cooperation, which could be a promising exploration direction in this area.

Possible extra costs: In our experiments, the AI utilized the labeling results together with the ground truth provided in exp 1 to form an assessment of its own and its human collaborators' capability, and then negotiated with humans to gradually reach an agreement in the subsequent process of forming the shared mental model. In reality, we could also design simple warm-up sessions which include frequent situations for human users and AI to develop initial capability assessments. Such a method is actually used by many existing mobile applications, which also try to collect the users' interests and preferences along the tutorial tasks when they start working with a new user, or obtain relevant information about the user from other platforms to improve the initial AI recommendations to human users [35]. For some extreme cases, assuming that a user does not have enough warm-up time to get along with the AI, we can also establish a built-in capability assessment of the general targeted user group and the AI itself, and continuously revise such assessment along with the updates of the shared mental model. We believe that such method is feasible based on our observation that the AI's capability assessment of users obtained in exp 1 is also relatively coarse, which could be similar for different users. However, in subsequent iterative negotiations with different users, it is able to lead different human-AI teams to reach a capability assessment and task assignment scheme close to their actual performance within about 30 minutes to 1 hour (the approximate duration of exp 3).

Another noteworthy aspect is the additional time loss of negotiation sessions we introduced, apart from cooperation working sessions. In Fig.7(a), when we computed the average time per task, we actually took into account the time loss of the negotiation sessions. Yet, the results show that the human-AI time efficiency is still better than the control group without shared mental model and negotiation sessions. It is probably due to the fact that human are more confident of the cooperation and their task assignment strategy, thus paying less attention to frequently questioning the AI's results or manually completing the labeling, since manually labelling the images is more time-consuming. In our experiments, we enter the negotiation sessions after every 30 pictures (tasks). In the future study or practical operations, the length of such gap can be adjusted according to the user's ability or availability for negotiation. For example, for autonomous driving tasks, we can try to guide the driver to complete such negotiation processes during the non-driving time or rest time to improve the cooperation experience.

5.2 Limitation and Future Work

We selected image labeling as our task to design the prototype system based on our shared mental model. We do not evaluate how people conceive a picture and categorize them from a cognitive behavioral science perspective, thus the number of chosen post-processing techniques is limited for simplification. Also, we do not consider AI with dynamic strategies to adjust algorithms well-trained for different scenarios to achieve the best accuracy. The AI algorithm used for our experiments does not change with different scenarios, and hence shows unsatisfying overall performance during the experiment we designed. However, such setting is in accordance with real-world occasions when merchandised AI agents have relatively stable performances but poor migration learning abilities, thus showing limited performance for complicated scenarios. Human and AI need to cooperate properly to achieve better efficacy.

The design of our mental model can also be improved if we take other variables into consideration. Such as whether to extend the list of conditions for simulating more complicated tasks, whether the confidence of human or AI would be treated as influential factors to align decisions during human-AI cooperation, whether to show the AI's numerical confidence in their choices, etc. We could also expand the complexity of negotiation methods other than bitwise operations that human-AI teams adapted for more efficient conversation. Other interaction channels such as eye movement and mouse movements, if could be collected during experiments and analyzed on the run, might also help to evaluate the engagement of the current user on the cooperation task. This could also take leaps forward towards more generalized and intelligent human-in-the-loop hybrid intelligent systems.

6 CONCLUSION

In this work, we clarified and designed the capability-aware shared mental model in task assignment based human-AI cooperation (TAHAC). It is a collaboratively negotiated, dynamically built model for grouping tasks based on human and AI's capability that could further be mapped into the task assignment. It is designed with 2 components: task grouping and negotiation. The task grouping takes on a basic representation unit as a tuple (m_i, s_i, a_i), which breaks down tasks' properties into sets of scenarios relating to task difficulties. The negotiation contains negotiation methods to dynamically merge the task grouping ideas raised by human and AI in an iterative form. We then implement a prototype system to evaluate capability-aware shared mental model via a collaborative image labelling task in a 3-phase user study. The results show that the introduced shared mental model could help with improving the accuracy and time efficiency of human-AI teams. It also helps human-AI teams come up with a task assignment plan close to their

real capability within few iterations. Users feel more confident while cooperating with AI, as well as form a better understanding of the AI's capability and capability difference between human and AI. Our designed capability-aware shared mental model shows potentials for effective task assignments of human-AI cooperation for various real-world tasks.

ACKNOWLEDGMENTS

We are grateful to Professor Xiaohong Guan for his kind support of this work and anonymous reviewers for their insightful comments. This work was funded in part by the National Key R&D Program of China (No. 2018AAA0101501).

REFERENCES

[1] Saleema Amershi, Daniel S. Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi T. Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019*, Stephen A. Brewster, Geraldine Fitzpatrick, Anna L. Cox, and Vassilis Kostakos (Eds.). ACM, Glasgow, Scotland, UK, 3. https://doi.org/10.1145/3290605.3300233

[2] Bahador Bahrami, Karsten Olsen, Peter E Latham, Andreas Roepstorff, Geraint Rees, and Chris D Frith. 2010. Optimally interacting minds. *Science* 329, 5995 (2010), 1081–1085.

[3] Victoria A Banks, Katherine L Plant, and Neville A Stanton. 2018. Driver error or designer error: Using the Perceptual Cycle Model to explore the circumstances surrounding the fatal Tesla crash on 7th May 2016. *Safety science* 108 (2018), 278–285.

[4] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, and Eric Horvitz. 2019. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. In *Proceedings of the Seventh AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2019, Stevenson, WA, USA, October 28-30, 2019*, Edith Law and Jennifer Wortman Vaughan (Eds.). AAAI Press, Stevenson, WA, USA, 2–11. https://ojs.aaai.org/index.php/HCOMP/article/view/5285

[5] Angie Boggust, Benjamin Hoover, Arvind Satyanarayan, and Hendrik Strobelt. 2022. Shared Interest: Measuring Human-AI Alignment to Identify Recurring Patterns in Model Behavior. In *CHI '22: CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April 2022 - 5 May 2022*, Simone D. J. Barbosa, Cliff Lampe, Caroline Appert, David A. Shamma, Steven Mark Drucker, Julie R. Williamson, and Koji Yatani (Eds.). ACM, New Orleans, LA, USA, 10:1–10:17. https://doi.org/10.1145/3491102.3501965

[6] Ángel Alexander Cabrera, Adam Perer, and Jason I Hong. 2023. Improving Human-AI Collaboration with Descriptions of AI Behavior. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 136:1–136:21.

[7] Tathagata Chakraborti, Sarath Sreedharan, Sachin Grover, and Subbarao Kambhampati. 2019. Plan explanations as model reconciliation–an empirical study. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, IEEE, Daegu, Korea (South), 258–266.

[8] Stevie Chancellor, Zhiyuan Lin, Erica L. Goodman, Stephanie Zerwas, and Munmun De Choudhury. 2016. Quantifying and Predicting Mental Illness Severity in Online Pro-Eating Disorder Communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, CSCW 2016, San Francisco, CA, USA, February 27 - March 2, 2016*, Darren Gergle, Meredith Ringel Morris, Pernille Bjørn, and Joseph A. Konstan (Eds.). ACM, San Francisco, CA, USA, 1169–1182.

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event (Proceedings of Machine Learning Research, Vol. 119)*. PMLR, Virtual Event, 1597–1607. http://proceedings.mlr.press/v119/chen20j.html

[10] Meghan Clark, Mark W Newman, and Prabal Dutta. 2017. Devices and data and agents, oh my: How smart home abstractions prime end-user mental models. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–26.

[11] Sharolyn Converse, JA Cannon-Bowers, and E Salas. 1993. Shared mental models in expert team decision making. *Individual and group decision making: Current issues* 221 (1993), 221–46.

[12] KJW Craik. 1943. The Nature of Explanation Cambridge University Press: Cambridge.

[13] Leslie A DeChurch and Jessica R Mesmer-Magnus. 2010. The cognitive underpinnings of effective teamwork: a meta-analysis. *Journal of applied psychology* 95, 1 (2010), 32.

[14] Maryam Banitalebi Dehkordi, Reda Mansy, Abolfazl Zaraki, Arpit Singh, and Rossitza Setchi. 2021. Explainability in human-robot teaming. *Procedia Computer Science* 192 (2021), 3487–3496.

[15] Michael Derntl, Renate Motschnig-Pitrik, and Kathrin Figl. 2006. Using teams, peer-and self evaluation in blended learning classes. In *Proceedings. Frontiers in Education. 36th Annual Conference*. IEEE, IEEE, San Diego, CA, USA, 15–20.

[16] David Eccles. 2010. The coordination of labour in sports teams. *International Review of Sport and Exercise Psychology* 3, 2 (2010), 154–170. https://doi.org/10.1080/1750984X.2010.519400 arXiv:https://doi.org/10.1080/1750984X.2010.519400

[17] Shirine El Zaatari, Mohamed Marei, Weidong Li, and Zahid Usman. 2019. Cobot programming for collaborative industrial tasks: An overview. *Robotics and Autonomous Systems* 116 (2019), 162–180.

[18] Jonnro Erasmus, Irene Vanderfeesten, Konstantinos Traganos, Ad Kleingeld, Paul Grefen, et al. 2018. A method to enable ability-based human resource allocation in business process management systems. In *IFIP Working Conference on the Practice of Enterprise Modeling*. Springer, Springer, Vienna, Austria, 37–52.

[19] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. 2016. Image Style Transfer Using Convolutional Neural Networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, Las Vegas, NV, USA, 2414–2423. https://doi.org/10.1109/CVPR.2016.265

[20] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. 2019. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, New Orleans, LA, USA, 22. https://openreview.net/forum?id=Bygh9j09KX

[21] Robert Geirhos, Carlos R. Medina Temme, Jonas Rauber, Heiko H. Schütt, Matthias Bethge, and Felix A. Wichmann. 2018. Generalisation in humans and deep neural networks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (Eds.). Curran Associates, Inc., Montréal, Canada, 7549–7561.

[22] Katy Ilonka Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz, Sarah Miller, David R. Millen, Murray Campbell, Sadhana Kumaravel, and Wei Zhang. 2020. Mental Models of AI Agents in a Cooperative Game Setting. In *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*, Regina Bernhaupt, Florian 'Floyd' Mueller, David Verweij, Josh Andres, Joanna McGrenere, Andy Cockburn, Ignacio Avellino, Alix Goguey, Pernille Bjøn, Shengdong Zhao, Briane Paul Samson, and Rafal Kocielnik (Eds.). ACM, Honolulu, HI, USA, 1–12. https://doi.org/10.1145/3313831.3376316

[23] Felix Gervits, Dean Thurston, Ravenna Thielstrom, Terry Fong, Quinn Pham, and Matthias Scheutz. 2020. Toward Genuine Robot Teammates: Improving Human-Robot Team Performance Using Robot Shared Mental Models. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems, AAMAS '20, Auckland, New Zealand, May 9-13, 2020*, Amal El Fallah Seghrouchni, Gita Sukthankar, Bo An, and Neil Yorke-Smith (Eds.). International Foundation for Autonomous Agents and Multiagent Systems, Auckland, New Zealand, 429–437. https://doi.org/10.5555/3398761.3398815

[24] Aaron Glick, Mackenzie Clayton, Nikola Angelov, and Jennifer Chang. 2022. Impact of explainable artificial intelligence assistance on clinical decision-making of novice dental clinicians. *JAMIA open* 5, 2 (2022), ooac031.

[25] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. XAI—Explainable artificial intelligence. *Science Robotics* 4, 37 (2019), 7120.

[26] Danula Eranjith Hettiachchi Mudiyanselage. 2021. *Task assignment using worker cognitive ability and context to improve data quality in crowdsourcing*. Ph. D. Dissertation. The University of Melbourne, Victoria, Australia.

[27] Andreas Holzinger. 2018. From machine learning to explainable AI. In *2018 world symposium on digital intelligence for systems and machines (DISA)*. IEEE, IEEE, Košice, Slovakia, 55–66.

[28] Andreas Holzinger, Anna Saranti, Christoph Molnar, Przemyslaw Biecek, and Wojciech Samek. 2020. Explainable AI methods-a brief overview. In *xxAI - Beyond Explainable AI - International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers (Lecture Notes in Computer Science, Vol. 13200)*, Andreas Holzinger, Randy Goebel, Ruth Fong, Taesup Moon, Klaus-Robert Müller, and Wojciech Samek (Eds.). Springer, Springer, Vienna, Austria, 13–38. https://doi.org/10.1007/978-3-031-04083-2_2

[29] Catholijn M. Jonker, M. Birna van Riemsdijk, and Bas Vermeulen. 2010. Shared Mental Models - A Conceptual Analysis. In *Coordination, Organizations, Institutions, and Norms in Agent Systems VI - COIN 2010 International Workshops, COIN@AAMAS 2010, Toronto, Canada, May 2010, COIN@MALLOW 2010, Lyon, France, August 2010, Revised Selected Papers (Lecture Notes in Computer Science, Vol. 6541)*, Marina De Vos, Nicoletta Fornara, Jeremy V. Pitt, and George A. Vouros (Eds.). Springer, Lyon, France, 132–151. https://doi.org/10.1007/978-3-642-21268-0_8

[30] Ece Kamar. 2016. Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (New York, New York, USA) *(IJCAI'16)*. AAAI Press, New York, NY, USA, 4070–4073.

[31] Ece Kamar, Severin Hacker, and Eric Horvitz. 2012. Combining human and machine intelligence in large-scale crowdsourcing. In *International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2012, Valencia, Spain, June 4-8, 2012 (3 Volumes)*, Wiebe van der Hoek, Lin Padgham, Vincent Conitzer, and Michael Winikoff (Eds.). IFAAMAS, Valencia, Spain, 467–474. http://dl.acm.org/citation.cfm?id=2343643

[32] H Kaur, A Williams, and WS Lasecki. 2019. Building shared mental models between humans and ai for effective collaboration. In *Proceedings of CHI 2019 Workshop on Where is the Human? Bridging the Gap Between AI and HCI, Glasgow, Scotland*. ACM, Glasgow, Scotland, 7.

[33] Matthew Kay, Tara Kola, Jessica R. Hullman, and Sean A. Munson. 2016. When (ish) is My Bus?: User-centered Visualizations of Uncertainty in Everyday, Mobile Predictive Systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, May 7-12, 2016*, Jofish Kaye, Allison Druin, Cliff Lampe, Dan Morris, and Juan Pablo Hourcade (Eds.). ACM, San Jose, CA, USA, 5092–5103. https://doi.org/10.1145/2858036.2858558

[34] William G Kennedy and J Gregory Trafton. 2007. *Using simulations to model shared mental models*. Technical Report. NAVAL RESEARCH LAB WASHINGTON DC CENTER FOR APPLIED RESEARCH IN ARTIFICIAL ....

[35] Julia Kiseleva, Alexander Tuzhilin, Jaap Kamps, Melanie J. I. Müller, Lucas Bernardi, Chad Davis, Ivan Kovacek, Mats Stafseng Einarsen, and Djoerd Hiemstra. 2016. Beyond Movie Recommendations: Solving the Continuous Cold Start Problem in E-commerce Recommendations. *CoRR* abs/1607.07904 (2016), 11. arXiv:1607.07904 http://arxiv.org/abs/1607.07904

[36] Lucy Van Kleunen and Stephen Voida. 2019. Challenges in supporting social practices around personal data for long-term mental health management. In *Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers, UbiComp/ISWC 2019 Adjunct, London, UK, September 9-13, 2019*, Robert Harle, Katayoun Farrahi, and Nicholas D. Lane (Eds.). ACM, London, UK, 944–948. https://doi.org/10.1145/3341162.3346273

[37] Richard Klimoski and Susan Mohammed. 1994. Team mental model: Construct or metaphor? *Journal of management* 20, 2 (1994), 403–437.

[38] Masaki Kobayashi, Kei Wakabayashi, and Atsuyuki Morishima. 2021. Human+AI Crowd Task Assignment Considering Result Quality Requirements. In *Proceedings of the Ninth AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2021, virtual, November 14-18, 2021*, Ece Kamar and Kurt Luther (Eds.). AAAI Press, Virtual, 97–107. https://ojs.aaai.org/index.php/HCOMP/article/view/18943

[39] Sonal Kothari, John H Phan, Todd H Stokes, and May D Wang. 2013. Pathology imaging informatics for quantitative analysis of whole-slide images. *Journal of the American Medical Informatics Association* 20, 6 (2013), 1099–1108.

[40] Walter S Lasecki. 2019. On facilitating human-computer interaction via hybrid intelligence systems. In *Proceedings of the 7th annual ACM Conference on Collective Intelligence. ACM.* ACM, Pittsburgh, USA, 1:1–1:5.

[41] Sunok Lee, Minji Cho, and Sangsu Lee. 2020. What If Conversational Agents Became Invisible? Comparing Users' Mental Models According to Physical Entity of AI Speaker. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–24.

[42] Lan Li, Tina Lassiter, Joohee Oh, and Min Kyung Lee. 2021. Algorithmic Hiring in Practice: Recruiter and HR Professional's Perspectives on AI Use in Hiring. In *AIES '21: AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event, USA, May 19-21, 2021*, Marion Fourcade, Benjamin Kuipers, Seth Lazar, and Deirdre K. Mulligan (Eds.). ACM, Virtual Event, USA, 166–176. https://doi.org/10.1145/3461702.3462531

[43] Yu Liangru Li Yi and Qiu Dong. 2020. Review of Cooperation Mode Between Human and Artificial Intelligence. *Journal of Intelligence* 39 (2020), 137–143.

[44] Claire Liang, Julia Proft, Erik Andersen, and Ross A. Knepper. 2019. Implicit Communication of Actionable Information in Human-AI teams. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019*, Stephen A. Brewster, Geraldine Fitzpatrick, Anna L. Cox, and Vassilis Kostakos (Eds.). ACM, Glasgow, Scotland, UK, 95. https://doi.org/10.1145/3290605.3300325

[45] Geert Litjens, Peter Bandi, Babak Ehteshami Bejnordi, Oscar Geessink, Maschenka Balkenhol, Peter Bult, Altuna Halilovic, Meyke Hermsen, Rob van de Loo, Rob Vogels, et al. 2018. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. *GigaScience* 7, 6 (2018), giy065.

[46] Alan Lundgard, Yiwei Yang, Maya L. Foster, and Walter S. Lasecki. 2018. Bolt: Instantaneous Crowdsourcing via Just-in-Time Training. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018*, Regan L. Mandryk, Mark Hancock, Mark Perry, and Anna L. Cox (Eds.). ACM, Montreal, QC, Canada, 467. https://doi.org/10.1145/3173574.3174041

[47] Lanssie Mingyue Ma, Terrence Fong, Mark J. Micire, Yunkyung Kim, and Karen M. Feigh. 2017. Human-Robot Teaming: Concepts and Components for Design. In *Field and Service Robotics, Results of the 11th International Conference, FSR 2017, Zurich, Switzerland, 12-15 September 2017 (Springer Proceedings in Advanced Robotics, Vol. 5)*, Marco Hutter and Roland Siegwart (Eds.). Springer, Zurich, Switzerland, 649–663. https://doi.org/10.1007/978-3-319-67361-5_42

[48] John E Mathieu, Tonia S Heffner, Gerald F Goodwin, Eduardo Salas, and Janis A Cannon-Bowers. 2000. The influence of shared mental models on team process and performance. *Journal of applied psychology* 85, 2 (2000), 273.

[49] Gerald Matthews, April Rose Panganiban, Jinchao Lin, Michael Long, and Michaela Schwing. 2021. Chapter 3 - Super-machines or sub-humans: Mental models and trust in intelligent autonomous systems. In *Trust in Human-Robot Interaction*, Chang S. Nam and Joseph B. Lyons (Eds.). Academic Press, Salt Lake City, UT, USA, 59–82. https://doi.org/10.1016/B978-0-12-819472-0.00003-4

[50] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[51] Susan Mohammed and Brad C Dumville. 2001. Team mental models in a team knowledge framework: Expanding theory and measurement across disciplinary boundaries. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior* 22, 2 (2001), 89–106.

[52] An T. Nguyen, Aditya Kharosekar, Saumyaa Krishnan, Siddhesh Krishnan, Elizabeth Tate, Byron C. Wallace, and Matthew Lease. 2018. Believe it or not: Designing a Human-AI Partnership for Mixed-Initiative Fact-Checking. In *The 31st Annual ACM Symposium on User Interface Software and Technology, UIST 2018, Berlin, Germany, October 14-17, 2018*, Patrick Baudisch, Albrecht Schmidt, and Andy Wilson (Eds.). ACM, Berlin, Germany, 189–199. https://doi.org/10.1145/3242587.3242666

[53] Stefanos Nikolaidis and Julie Shah. 2012. Human-robot teaming using shared mental models. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction 2012 Workshop on Human-Agent-Robot Teamwork, Boston, Massachusetts, USA*. Association for Computing Machinery, Boston, Massachusetts, USA, 6.

[54] Besmira Nushi, Ece Kamar, and Eric Horvitz. 2018. Towards Accountable AI: Hybrid Human-Machine Analyses for Characterizing System Failure. In *Proceedings of the Sixth AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2018, Zürich, Switzerland, July 5-8, 2018*, Yiling Chen and Gabriella Kazai (Eds.). AAAI Press, Zürich, Switzerland, 126–135. https://aaai.org/ocs/index.php/HCOMP/HCOMP18/paper/view/17930

[55] Mike Oaksford and Nick Chater. 2001. The probabilistic approach to human reasoning. *Trends in cognitive sciences* 5, 8 (2001), 349–357.

[56] Kristin E. Oleson, Deborah R. Billings, Vivien Kocsis, Jessie Y. C. Chen, and Peter A. Hancock. 2011. Antecedents of trust in human-robot collaborations. In *IEEE International Inter-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support, CogSIMA 2011, Miami, FL, USA, February 21-24, 2011*, Gabriel Jakobson, Mica R. Endsley, and Mitch Kokar (Eds.). IEEE, Miami, FL, USA, 175–178. https://doi.org/10.1109/COGSIMA.2011.5753439

[57] Scott Ososky. 2013. *Influence of task-role mental models on human interpretation of robot motion behavior*. Ph. D. Dissertation. University of Central Florida, Orlando, Florida, USA. Advisor(s) Florian G. Jentsch.

[58] Scott Ososky, David Schuster, Florian Jentsch, Stephen Fiore, Randall Shumaker, Christian Lebiere, Unmesh Kurup, Jean Oh, and Anthony Stentz. 2012. The importance of shared mental models and shared situation awareness for transforming robots from tools to teammates. In *Unmanned systems technology XIV*, Vol. 8387. International Society for Optics and Photonics, International Society for Optics and Photonics, Baltimore, Maryland, United States, 838710.

[59] Nicholas Paul and ChanJin Chung. 2018. Application of HDR algorithms to solve direct sunlight problems when autonomous vehicles using machine vision systems are driving into sun. *Computers in Industry* 98 (2018), 192–196.

[60] Erasmo Purificato, Flavio Lorenzo, Francesca Fallucchi, and Ernesto William De Luca. 2022. The Use of Responsible Artificial Intelligence Techniques in the Context of Loan Approval Processes. *International Journal of Human–Computer Interaction* 0, 0 (2022), 1–20. https://doi.org/10.1080/10447318.2022.2081284

[61] Katyanna Quach. 2020. Watch an oblivious Tesla Model 3 smash into an overturned truck on a highway 'while under autopilot'. https://www.theregister.com/2020/06/02/tesla_car_crash/

[62] Yosef S Razin, Jack Gale, Jiaojiao Fan, Jaznae' Smith, and Karen M Feigh. 2021. Watch For Failing Objects: What Inappropriate Compliance Reveals About Shared Mental Models In Autonomous Cars. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 65. SAGE Publications Sage CA: Los Angeles, CA, Los Angeles, CA, USA, 643–647.

[63] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "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* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[64] Peter J Robe. 2021. *Designing a Pair Programming Conversational Agent*. Ph. D. Dissertation. The University of Tulsa.

[65] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C.

Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115, 3 (2015), 211–252. https://doi.org/10.1007/s11263-015-0816-y

[66] Mike Schaekermann, Graeme Beaton, Elaheh Sanoubari, Andrew Lim, Kate Larson, and Edith Law. 2020. Ambiguity-aware AI Assistants for Medical Data Analysis. In *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*, Regina Bernhaupt, Florian 'Floyd' Mueller, David Verweij, Josh Andres, Joanna McGrenere, Andy Cockburn, Ignacio Avellino, Alix Goguey, Pernille Bjøn, Shengdong Zhao, Briane Paul Samson, and Rafal Kocielnik (Eds.). ACM, Honolulu, HI, USA, 1–14. https://doi.org/10.1145/3313831.3376506

[67] Mike Schaekermann, Carrie J. Cai, Abigail E. Huang, and Rory Sayres. 2020. Expert Discussions Improve Comprehension of Difficult Cases in Medical Image Assessment. In *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*, Regina Bernhaupt, Florian 'Floyd' Mueller, David Verweij, Josh Andres, Joanna McGrenere, Andy Cockburn, Ignacio Avellino, Alix Goguey, Pernille Bjøn, Shengdong Zhao, Briane Paul Samson, and Rafal Kocielnik (Eds.). ACM, Honolulu, HI, USA, 1–13. https://doi.org/10.1145/3313831.3376290

[68] Matthias Scheutz. 2013. Computational Mechanisms for Mental Models in Human-Robot Interaction. In *Virtual Augmented and Mixed Reality. Designing and Developing Augmented and Virtual Environments - 5th International Conference, VAMR 2013, Held as Part of HCI International 2013, Las Vegas, NV, USA, July 21-26, 2013, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 8021)*, Randall Shumaker (Ed.). Springer, Las Vegas, NV, USA, 304–312. https://doi.org/10.1007/978-3-642-39405-8_34

[69] Francesco Secci and Andrea Ceccarelli. 2020. On failures of RGB cameras and their effects in autonomous driving applications. In *31st IEEE International Symposium on Software Reliability Engineering, ISSRE 2020, Coimbra, Portugal, October 12-15, 2020*, Marco Vieira, Henrique Madeira, Nuno Antunes, and Zheng Zheng (Eds.). IEEE, Coimbra, Portugal, 13–24. https://doi.org/10.1109/ISSRE5003.2020.00011

[70] Peter Selinger. 2003. Potrace–Transforming bitmaps into vector graphics. [EB/OL]. http://potrace.sourceforge.net/#license Accessed July 22, 2021.

[71] Donghee Shin. 2021. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies* 146 (2021), 102551.

[72] Ben Shneiderman. 2022. Human-Centered AI: Ensuring Human Control While Increasing Automation. In *Proceedings of the 5th Workshop on Human Factors in Hypertext* (Barcelona, Spain) *(HUMAN '22)*. Association for Computing Machinery, New York, NY, USA, Article 1, 2 pages. https://doi.org/10.1145/3538882.3542790

[73] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). OpenReview.net, San Diego, CA, USA, 14. http://arxiv.org/abs/1409.1556

[74] Kimberly A Smith-Jentsch, John E Mathieu, and Kurt Kraiger. 2005. Investigating linear and interactive effects of shared mental models on safety and efficiency in a field setting. *Journal of applied psychology* 90, 3 (2005), 523.

[75] Kihyuk Sohn, Chun-Liang Li, Jinsung Yoon, Minho Jin, and Tomas Pfister. 2021. Learning and Evaluating Representations for Deep One-Class Classification. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, Virtual, 32. https://openreview.net/forum?id=HCSgyPUfeDj

[76] Jean Y Song, Raymond Fok, Juho Kim, and Walter S Lasecki. 2019. FourEyes: Leveraging Tool Diversity as a Means to Improve Aggregate Accuracy in Crowdsourcing. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 10, 1 (2019), 1–30.

[77] Jean Y. Song, Stephan J. Lemmer, Michael Xieyang Liu, Shiyan Yan, Juho Kim, Jason J. Corso, and Walter S. Lasecki. 2019. Popup: reconstructing 3D video using particle filtering to aggregate crowd responses. In *Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI 2019, Marina del Ray, CA, USA, March 17-20, 2019*, Wai-Tat Fu, Shimei Pan, Oliver Brdiczka, Polo Chau, and Gaelle Calvary (Eds.). ACM, Marina del Ray, CA, USA, 558–569. https://doi.org/10.1145/3301275.3302305

[78] Rooji Sugathan, Shaji Khan, Dinesh Mirchandani, and Ashok Subramanian. 2020. *System Usage: A Shared Mental Model Perspective.* Ph. D. Dissertation. University of Missouri - Saint Louis. Advisor(s) Vicki, Sauter,. AAI27960557.

[79] Piet Van den Bossche, Wim Gijselaers, Mien Segers, Geert Woltjer, and Paul Kirschner. 2011. Team learning: building shared mental models. *Instructional Science* 39, 3 (2011), 283–301.

[80] Jeroen Van der Laak, Geert Litjens, and Francesco Ciompi. 2021. Deep learning in histopathology: the path to the clinic. *Nature medicine* 27, 5 (2021), 775–784.

[81] Lev Velykoivanenko, Kavous Salehzadeh Niksirat, Noé Zufferey, Mathias Humbert, Kévin Huguenin, and Mauro Cherubini. 2021. Are Those Steps Worth Your Privacy? Fitness-Tracker Users' Perceptions of Privacy and Utility. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (2021), 1–41.

[82] Zijie J Wang, Robert Turko, Omar Shaikh, Haekyu Park, Nilaksh Das, Fred Hohman, Minsuk Kahng, and Duen Horng Polo Chau. 2020. CNN explainer: Learning convolutional neural networks with interactive visualization. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2020), 1396–1406.

[83] Bryan Wilder, Eric Horvitz, and Ece Kamar. 2020. Learning to Complement Humans. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, Christian Bessiere (Ed.). International Joint Conferences on Artificial Intelligence Organization, Virtual, 1526–1533. https://doi.org/10.24963/ijcai.2020/212 Main track.

[84] Ming Yin, Jennifer Wortman Vaughan, and Hanna M. Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019*, Stephen A. Brewster, Geraldine Fitzpatrick, Anna L. Cox, and Vassilis Kostakos (Eds.). ACM, Glasgow, Scotland, UK, 279. https://doi.org/10.1145/3290605.3300509

[85] Luyao Yuan, Xiaofeng Gao, Zilong Zheng, Mark Edmonds, Ying Nian Wu, Federico Rossano, Hongjing Lu, Yixin Zhu, and Song-Chun Zhu. 2022. In situ bidirectional human-robot value alignment. *Science robotics* 7, 68 (2022), eabm4183.

[86] Nor'ain Mohd Yusoff and Siti Salwah Salim. 2020. Shared Mental Model Processing in Visualization Technologies: A Review of Fundamental Concepts and a Guide to Future Research in Human-Computer Interaction. In *Engineering Psychology and Cognitive Ergonomics. Mental Workload, Human Physiology, and Human Energy - 17th International Conference, EPCE 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19-24, 2020, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 12186)*, Don Harris and Wen-Chin Li (Eds.). Springer, Copenhagen, Denmark, 238–256. https://doi.org/10.1007/978-3-030-49044-7_20

[87] Matthew D. Zeiler and Rob Fergus. 2014. Visualizing and Understanding Convolutional Networks. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 8689)*, David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer, Zurich, Switzerland, 818–833. https://doi.org/10.1007/978-3-319-10590-1_53

[88] Rui Zhang, Nathan J McNeese, Guo Freeman, and Geoff Musick. 2021. " An Ideal Human" Expectations of AI Teammates in Human-AI Teaming. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–25.