Six Challenges for Human-AI Co-learning

Karel van den Bosch[1(✉)], Tjeerd Schoonderwoerd[1], Romy Blankendaal[1], and Mark Neerincx[1,2]

[1] TNO, Kampweg 55, 3769 DE Soesterberg, The Netherlands
{karel.vandenbosch,tjeerd.schoonderwoerd,romy.blankendaal,
mark.neerincx}@tno.nl
[2] Delft University of Technology, Van Mourik Broekmanweg 6,
2628 XE Delft, The Netherlands

Abstract. The increasing use of ever-smarter AI-technology is changing the way individuals and teams learn and perform their tasks. In hybrid teams, people collaborate with artificially intelligent partners. To utilize the different strengths and weaknesses of human and artificial intelligence, a hybrid team should be designed upon the principles that foster successful human-machine learning and cooperation. The implementation of the identified principles sets a number of challenges. Machine agents should, just like humans, have mental models that contain information about the task context, their own role (self-awareness), and the role of others (theory of mind). Furthermore, agents should be able to express and clarify their mental states to partners. In this paper we identify six challenges for humans and machines to collaborate in an adaptive, dynamic and personalized fashion. Implications for research are discussed.

Keywords: Co-active learning · Human-agent teaming · Hybrid teams · Theory of mind · Explainable AI · Mental model

1 Introduction

The literature on teams (e.g., [48,52]) has produced knowledge on how to design a training environment and the operational environment to ensure that a team of experts is also an expert team [47]. Now, with the introduction of advanced technology, people also have to form effective teams with artificially intelligent partners. The principles derived from studies on the effectiveness of human-human teams are valuable for designing human-technology teams, but there are also differences between human intelligence and Artificial Intelligence (from now on: AI) that must be taken into account. Modern AI-applications acquire knowledge about their domain and tasks by establishing correlations and patterns in the large sets of data they collect about their environment. It then uses this knowledge to solve new problems. When the environment provides sufficient data, the algorithm can become very successful (e.g., for example, recognizing

R. A. Sottilare and J. Schwarz (Eds.): HCII 2019, LNCS 11597, pp. 572–589, 2019.
https://doi.org/10.1007/978-3-030-22341-0_45

cancerous tissue in MR-images [62]). However, the intelligence of such applications remain within the boundaries of the trained task. If these are narrow and well-defined, then AI is doing well. However, when the task context imposes a rich and an a priori unknown variety of conditions (wide system boundaries), then the problem-solving intelligence of AI drops dramatically [3]. Where AI still falls short is thinking in the abstract, in applying common sense, and in transferring knowledge from one area to another [7]. Thus, humans and AI each have their strengths and weaknesses. Humans, for example, are poor at storing and processing information, unlike AI. However, humans can make abstract decisions based on intuition, common sense, and scarce information [29]. Rather than acting as separate and equal entities, humans and AI should collaborate in a coordinated fashion to unlock the strengths of a heterogeneous team. It is believed that this needs to develop iteratively by interaction between partners [4,19,27]. This paper discusses the challenges for developing systems that enable humans and artificially intelligent technology to jointly learn and work together, adaptively and effectively.

1.1 Hybrid Teams

A hybrid team is a team of multiple agents that interdependently work together, and where agents can be either humans or machines. The cooperation of humans and machines sets new demands as the nature of intelligence is different between agents [3]. One demand is that the conditions must be created in which all agents come to recognize and acknowledge their respective capabilities. This may apply to a single human-machine combination, but it may also concern a team of multiple human-machine combinations. Another demand is that team members should have a shared understanding of how to exploit their complementary strengths to the benefit of the team. How team members should adapt to form an effective team varies from occasion to occasion. It is dependent upon many factors, like for example the specific demands of the context, the capabilities and preferences of the other team members, and many other variables. Learning how to adapt always take place, with every new performance of a team. Each training and each operation provides opportunities for team members to develop their skills, to refine their understanding of their own role within the team, and to deepen their knowledge of the other team members' roles, capabilities, and preferences. A further demand is that the members of a hybrid team should be able to use the progressive insights of its members to formalize and tune the work agreements. Figure 1 shows a representation of a hybrid team.

The green inner area of Fig. 1 shows a team consisting of four human-machine unit. One human-machine unit is shown enlarged for explication purposes. The human and the machine both have, develop, and maintain a mental model (shown in the lower two clouds). The mental model of the human involves knowledge about the task, a concept of its own role in the team, and expectations about the contributions of other team members. The mental model of a machine is likely to be much less elaborated, involving specific knowledge about the task to be performed by the machine, and some aspects of the context.

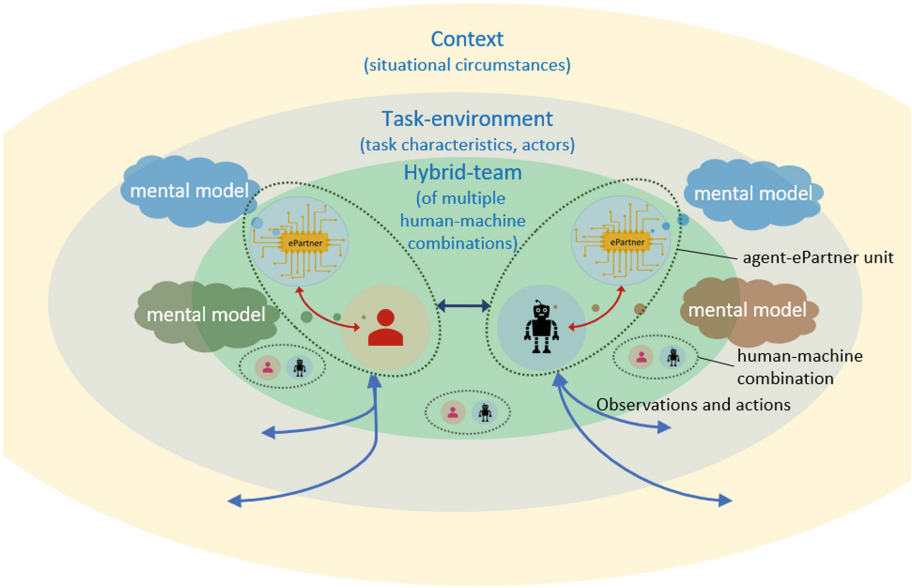


Fig. 1. Human-machine cooperation in a hybrid team (Color figure online)

Both the human and the machine have an ePartner, an AI-based agent [5,40]. The purpose of an ePartner is to assist its user to act as a good team member. An agent (either machine or human) and its ePartner form a unit (hence the dotted ellipses). Indirectly, by assisting its user, the ePartner supports the team as a whole. An ePartner collects and processes information about the task (grey middle zone) and the context (outer yellow zone). It also collects and processes information about its user, about the partner of its user, as well as of the other human-machine units of the team (green inner zone). The ePartners use this information to construct and maintain an elaborated mental model (shown in the two upper blue clouds) containing a representation of its user (a 'self'-model), as well as a representation of the perspective of the partner-agent (a theory of mind model). Through these mental models the ePartners develop an understanding of the task, users, and team. Based on this understanding, the ePartner can initiate various support actions.

A hybrid team consists of agents. An agent is an entity that is autonomous, intentional, social, reactive, and proactive [61]. So a human is an agent. A machine can also be an agent, but only if it meets the criteria above. For instance, a robot arm that mechanically performs some kind of action is not considered an agent, even though its actions may be valuable to the team. In the hybrid team outlined in Fig. 1, we have machines in mind that are more or less intelligent agents. A machine's mental model is typically targeted at the intelligence needed for acting adequately in a bounded variety of task conditions. It generally does not include the ability to acknowledge the needs of its partner, or of other members of the team. Thus a mental model of the machine supports task behavior, not

team behavior [51]. However, in a machine-ePartner unit (right dotted ellipse), the ePartner-agent is able to develop a mental model that covers the needs of others; not only of its machine-partner, but also of other agents in the team. This enables this ePartner to initiate supportive actions (e.g., informing its machine that a task has already been done by others in the team; informing the ePartner of the human partner that the machine's battery is about empty). In a human-ePartner unit (left dotted ellipse), both the human as well as its ePartner-agent develop a mental model of the task and of the team. However, the mental models of these agents are not the same. The human's ePartner-agent can, for example, receive information from the ePartner of the machine about the status of its task work (e.g., the remaining battery power of the machine). It can also determine conditions of its partner (e.g., fatigued; high-work load) that the human may self not be aware of [16]. Again, this enables the ePartner to initiate supportive actions (e.g., issue warning to human partner; request other agents in the team to take over tasks).

This envisioned cooperation between humans and machines in a hybrid team needs to develop through interaction and feedback during learning and operations, enabling all agents to acquire implicit and explicit knowledge about themselves and about their partners. Implicit knowledge about the partner is, for example, intuitively knowing how the partner will respond to a particular situation (often without realizing why). This is called 'tacit knowledge' [46], as it often cannot be adequately articulated. Explicit knowledge is, for example, knowing what the partner is likely to achieve, and accordingly, how it will act. Explicit knowledge is often obtained by deduction, logic, and reasoning [10].

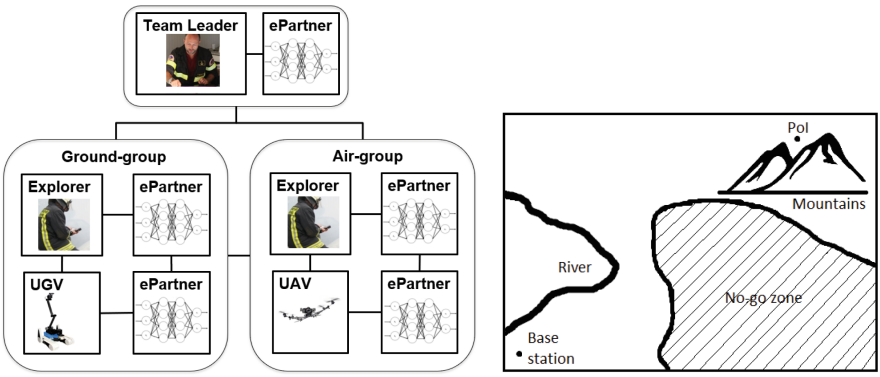
The next chapter presents a use case of human-AI co-learning in hybrid teams, relating it to the literature for principles of successful development of human-AI partnerships. These principles are used to define the challenges for establishing human-AI Co-learning in Sect. 3. The final chapter discusses the implications for research.

2 Co-learning in Hybrid Teams

The increasing use of ever-smarter AI technology is changing the way individuals and teams perform their tasks. Designing the models for successful hybrid teams should be based upon the principles that foster the cooperation between units consisting of human-machine combinations, and that promote the collaboration of multiple human-machine combinations at the team level (see Fig. 1). This section proposes a set of principles for human-AI co-learning, derived from the literature on human-machine interaction, human-agent teaming, and teamwork in general. It starts with a general use case to illustrate the co-learning process.

2.1 Use Case

Figure 2a presents an overview of a Human-ePartner-Robot-Team (HeRT) at a disaster scene of our use case (inspired by the TRADR use cases for robot-assisted



(a) Human-ePartner-Robot-Team (b) Example scenario

Fig. 2. HeRT team and scenario map (scene)

disaster response; [14,28]). In this team, all agents have sensors for monitoring the environment (e.g., to identify human beings, passageways, objects) and their own states (e.g., health and location). However, machine agents will only have limited state-sensing capability. To support collaboration, humans and ePartners are also equipped with sensors that assess states of other agents (such as workload; [16]). There is a shared knowledge base; policies define the obligations, permissions and prohibitions for knowledge exchange (e.g., as adjustable work agreements [38]).

When approaching the disaster scene, the Team Leader (TL) assesses the situation, selects the first Point-of-Interests (PoI) to explore, and estimates the corresponding priorities. Based on previous missions, the ePartner of the Team Leader, i.e., ePartner(TL), proposes task allocations and work agreements for the team. Figure 2b gives an overview of the task context. The PoI is in a valley in between mountains. Victims might be found there, but the area is dangerous for humans due to the possible presence of toxic gases. The ePartners initiate the operation by issuing work agreements among the groups and units, i.e., for notifying progress, agent states, and environmental events. A first example is that the TL will be notified about the progress of (all) groups, when there is a (i) deviation of the plan, (ii) change of agent's state, or (iii) unforeseen critical event. Second, specific for the Air-group, there is the agreement that the TL-unit will get regular updates ("situation reports", provided by the ePartner of air-group's Explorer, i.e., ePartner(E-air)) with the overview pictures of the UAV (so that TL will maintain a general overview, and can immediately help the less experienced Air-group when needed). Third, specific between the Ground-group and the Air-group, there is the agreement that the other group is notified when a new obstacle is detected in the planned navigation routes.

Following the plan, the Explorer(air) initiates the first (high-priority) task: The UAV has to explore the area between the base station and PoI to assess its accessibility for UGV navigation. In parallel, Explorer(ground) initiates the first (high-priority) task: navigation of UGV to PoI, to gather information about

the environment during the navigation and, subsequently, at PoI. Based on the available information of the environment, the UGV calculates the best navigation route and starts navigating. The Air-group identifies a blockade of the planned route and ePartner(E-air) notifies the Ground-group. The UGV changes its route and continues; ePartner(UGV) provides an explanation; ePartner(TL) notifies the TL about the changed route plan with the explanation (and the information that the time of arrival at PoI is extended).

In the meantime, the Air-group (i) is processing a large amount of environmental data with inconclusive outcomes, (ii) has to anticipate for a required battery change, and (iii) is notified that storm and rain are approaching. The ePartner(E-air) identifies a "cognitive lock-up" in the data-processing task of its partner, draws her attention to the battery level and weather forecast, and notifies unit(TL). The TL assesses the adapted task plan, UAV's battery level and the weather forecast, and determines that the UAV can stay in air till the UGV approaches the PoI.

After the mission, all agents participate in a debriefing session. The ePartner(E-air) points to the cognitive-lock-up event, and explains its assessment. The TL refines the explanation, enhancing ePartners' knowledge base. Explorer(air) understands what happened and selects training scenarios to practice this type of situations in virtual reality.

2.2 Principles of Human-AI Co-learning

Research in human-machine interaction provides useful models and methods for the required communication in the envisioned use case of Sect. 2.1, such as chat bots [11], virtual assistants [16,54], and personal teaching agents [25,55]. ePartners should tailor their communication to the specific characteristics of their human partner (e.g., preferences, experiences, mental state), the team (e.g., roles, work procedures, communication protocols) and the context (e.g., movement, noise, time pressure). However, collaboration and collaborative learning are not driven by explicit demarcated communicative acts only. A joint task performance of human and a machine agent requires that their social, cognitive, affective and physical behaviors are harmonized for the work processes. For establishing such harmonization, we identify a number of important principles: OPED (observability, predictability, explainability & directability), trust generation & calibration, self-awareness & theory of mind, lifelong learning on the job, and teams learning from teams.

Observability, Predictability, Explainability and Directability
Joint task performance requires that the agents deal with interdependencies: the coordinated adaptation of task performance of humans and machines to optimize their performance as a team [19,42]. Johnson et al. define three requirements for successful interdependent collaboration: Observability, Predictability and Directability [19]. In addition, Explainability has been identified as an important prerequisite for collaboration and learning (e.g., [12,41]).

Observability implies that the human agent and the machine agent are informed of their own actions, each other's actions, and the status of their role and progress in the task. In a human-machine partnership this requires that the state of a machine agent should be observable to a human partner, and the machine-agent should be informed about the human's status from explicit and implicit behavioral cues. The use case of Sect. 2.1 provides several example work agreements for establishing observability within a team (e.g., on agent's state, like robot's battery level and explorer's stress level).

Predictability means that actions of a team member are -to some extent- predictable, so that team members can understand it, and anticipate to it. The use case shows, for example, the processing of prediction information within the team on robot state (battery level), weather and reaching the PoI, to decide on the UAV's route.

Explainability is needed in circumstances where partners desire a clarification of each other's behavior. One way of achieving this is by requesting explanations. In order to generate an explanation that fits the objective of the requesting agent, partners should have the capabilities to diagnose the state of the other agent (related to observability), and the partner's intention of the request (related to predictability). In the use case of Sect. 2.1, for example, the ePartner of the UGV provides an explanation of the changed route towards the PoI.

Directability refers to the property of agents to take over and delegate tasks, both reactively and pro - actively. In the use case, for example, the TL takes over part of the task of explorer(air), when she is in a "lock-up".

Trust Generation and Calibration

The research community has not (yet) provided a unified definition of trust, but it is commonly recognized that trust is as a psychological state that is influenced by the complex interrelations between expectations, intentions and dispositions [9]. For now, we will use Mayer's trust definition: "The willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party" [36] (p. 712). Trust development is a continuous process in teamwork, involving trust establishment and adjustment based on team-members' experiences concerning each other's performances and the overall team performance. In teamwork, three processes should be considered: (i) interpersonal trust between members, (ii) collective trust at the team level, and (iii) the cross-level influences and dynamics [9]. It should be noted that high levels of team trust may have negative consequences, like the pressure to conform to group's norms in "groupthink" [18]. Adequate trust calibration is crucial to establish appropriate attitudes and performances in teamwork.

In the use case of Sect. 2.1, for example, the TL developed a higher level of trust for the (experienced) Ground group than for the (less experienced) Air group. Based on the low level of trust, a specific work agreement was made for the last group: ePartner(E-air) provides regular updates with overview pictures to unit(TL), so that the TL can immediately help the Air-group when needed.

Self-awareness and Theory of Mind
In a well-functioning team, the team-members learn to perform the tasks, how their tasks relate to those of the other(s), and how to manage their tasks. They develop "self-awareness" of their own state and role in the partnership, "self-management capabilities" and a "Theory of Mind" (i.e., knowledge of the other, [45]). Current AI developments are enhancing machine's capabilities on the "self-awareness & management" [59]. However, developing a Theory of Mind also proves to be crucial for effective human-human teamwork [35] and human-machine teamwork [24,30,43,53,60]. Furthermore, the self-awareness, self-management and Theory of Mind can develop at four levels: agent, unit, group and team. The ePartners aim to enhance this by sensing, modeling, activating and sharing the relevant information (see Fig. 1). With the capabilities to form Theories of Mind, humans and ePartners develop the capability to maintain common ground, thereby meeting the challenge of Klein and colleagues [26] for successful joint activity. In the use case of Sect. 2.1, ePartner(E-air) detected a "cognitive lock-up" of its partner, sharing it (i) with the TL-unit (team level) to ensure an effective air-group task performance, and (ii) with its partner ("self-awareness at unit level") for experienced-based learning.

Lifelong Learning on the Job
Appropriate experience sharing will help teams to learn from their practices and improve their adaptive capabilities. For example, team reflections can make "tacit" knowledge explicit in a systematic way in such a way that the team can better cope with similar situations in the future (i.e., team's resilience increases, [49]). The ePartner will support this process by (i) providing the "episodic memory" with the features that affect performance and resilience, and (ii) the procedures to reflect on these episodes [16]. One way to do this, is to share experiences, and to reflect upon these experiences [58], for example by engaging in an After Action Review [39].

The previous principle ("Self-awareness and Theory of Mind"), already referred to the learning of Explorer(air) in the use case by recalling the "cognitive lock-up" episode. In addition, the TL-unit will learn from this episode about the effectiveness of its back-up behavior (i.e., enhancing team's resilience).

Teams Learning from Teams
A learning organization requires that team experiences and knowledge are shared with other teams continuously (cf., [6]). Concerning this capability, ePartners will provide excellent support: Their knowledge-base can be, almost instantaneously and completely, shared with all the other ePartners. This way, an evolving library of constructive and destructive team patterns can be build and shared [57]. Subsequently, the ePartners can help to identify such patterns when they appear with the corresponding supporting or mitigating strategies. For the use case of Sect. 2.1, for example, the set of work agreements that proved to be effective will be shared by all teams.

3 Challenges for Developing Hybrid Team Agents

The previous chapter discussed the principles for human-AI co-learning from a team perspective. In this chapter we address the implementation of these principles: the challenges of creating learning human-AI partnerships, the constituting elements of a hybrid team.

An important prerequisite for effective task and team performance is that humans and machines become aware of each other's knowledge, skills, capabilities, goals, and intentions. Humans store and structure such information in their brain in the form of mental models [8,21]. Mental models can be regarded as personal and subjective interpretations of what something is, and how something works in the real world. Humans use their mental models to explain and predict the world around them, for example interpreting the behavior of others. In fact, it has been demonstrated that a mental model of the environment, including information about the task and knowledge of other agents, is required for efficient cooperation between humans in a team [35]. We argue that if a team consists of humans and machines, machine agents need to be initiated with a basic model of the task context, their own role, and the role of others. Furthermore, they need to be able to learn from experiences and feedback; to refine and adjust their mental models. Not all knowledge and functions need to reside in the individual agents; agents are able to share information, thus creating a kind of "team cloud" database. We identify the following six challenges to achieve effective human-machine team collaboration:

1. Agents of a hybrid team should have, develop, and refine a shared vocabulary of concepts and relations *(taxonomy model)*
2. Agents of a hybrid team should have access to a shared set of work agreements and interdependencies. This include agreements on how agents can dynamically update this as a result of learning (*team model*)
3. An agent should have, develop, and refine a mental model containing knowledge about the regularities between task conditions, actions and outcomes (*task model*)
4. An agent should have, develop, and refine a mental model containing knowledge about itself, including its needs, goals, values, capabilities, resources, plans, and emotions (*self-model*)
5. An agent should have, develop, and refine a mental model containing knowledge of other agent's needs, goals, values, capabilities, resources, plans, and emotions (*theory-of-mind model*)
6. An agent should have the functionalities, instruction, and training to communicate and explain experiences to other agents (*communication model*)

Challenges concerning the contents of agents' mental models are discussed in Sect. 3.1. The mental models of agents should not constitute a fixed representation of the world, but a dynamic one. The models' contents need to be constantly refined and adjusted, as a result of learning from experiences. This raises the question how machine agents should restructure their mental models in order to assimilate and represent newly acquired knowledge. Such representation

challenges are discussed in Sect. 3.2. A mental model is functional in the sense that it helps the agent to determine and tune its behavior and to develop an approach for solving a problem. At best, an agent's operations may be experienced as logical, plausible, or understandable by other agents. However, sometimes they lead to surprise or incomprehension. Establishing a flexible and resilient hybrid team requires mechanisms that enable agents to resolve misconceptions, ambiguities and inconsistencies. These challenges are discussed in Sect. 3.3.

3.1 Components of Mental Models

Conceptually, we distinguish between three types of integrated knowledge in a mental model of a hybrid team agent: knowledge about the task and context; knowledge about oneself; and knowledge about the partner.

Knowledge about the task and context: through instruction and experience, an agent accumulates its knowledge about about the regularities between task conditions, actions and outcomes. The agent should be able to expand its task model with the acquired knowledge (challenge 3). The agent may or may not be aware of its knowledge. Some of the relationships may be formally coded in the mental model (e.g., the task condition of seeing a 'stop' sign, triggers the act of stopping the car, leading to the outcome of a safe crossing of the intersection). An agent's mental model should also contain strategies for conducting a task. Formal knowledge about relationships and strategies can be easily communicated to other agents. In addition, agents may also have implicit knowledge of regularities in their mental model. For example, when having to make a right turn, a driver agent uses subtle environmental cues to apply the forces to the steering wheel and gas pedal that produce an adequate bend. The implicit nature of such knowledge, also called 'tacit knowledge' [44], makes it hard to articulate it, and thus to communicate it with other agents.

Knowledge about oneself: the mental model of a hybrid team agent should contain information about its own needs, goals, values, capabilities, resources, plans, and emotions (challenge 4). This enables the agent to be self-aware, an essential principle for self-management, as well as for alignment and adaptation in a team. An agent's self-knowledge should be adjustable under influence of interactions, experiences and feedback.

Knowledge about other(s): Agents should also be able to construct models of other agents (challenge 5), a theory of mind. The agent should have the metacognitive ability to attribute mental capacities and states to others [45], such as their assumed motivations, beliefs, values, goals, and aspects of personality. Furthermore, this theory-of-mind model should also include information about how the other agent thinks about its partner (i.e., "what could the other agent know about my knowledge, beliefs, values, and emotions?").

Another challenge is that an agent should be able to retrieve and connect information from the different sources of knowledge (challenge 1) so that the agent can detect and understand interdependencies within the team (challenge 2). It allows the agent to infer, for example, that a team agent may be too fatigued to carry out its task, and to offer assistance to this team agent.

3.2 Representational Challenges for Mental Models

As argued in Sect. 3.1, agents should be able to develop a mental model consisting of different types of information, like observations or factual information in the task environment (e.g., whether something or someone is present or not), known or perceived relationships between events and actions (e.g., "if I see fire, and I press this button, then an alarm will sound"), and assumptions (e.g., "if my partner is very busy, then he is more likely to ignore my request"). In a task domain, there often exist many relationships between different types of information. An agent's mental model should be able to represent all these, and the representations should allow the agent to make connective associations between them. The sections below discuss the challenges associated with this requirement.

Hybrid AI

The literature reports a variety of models that can represent an individual's performance and psychological states, such as emotion, trust, stress, memory, and theory of mind (see [31] for an overview). A symbolic approach, for example, is based on knowledge and rules and works best in well-defined problems. An advantage of symbolic models is that they are understandable to people. Another approach is to represent knowledge as a network of nodes and associations (e.g., [50]). This data-driven, sub-symbolic approach to modeling is suited for ill-defined problem environments. However, a disadvantage is that knowledge is distributed throughout the network, and is therefore non-transparent for humans.

It has been advocated to combine both approaches, for example as shown in Fig. 3. This is called hybrid AI [1,32,41,56]. Interestingly, human thinking is also considered to be the result of a combination of implicit intuitive knowledge (cf. sub-symbolic), and explicit, conscious reasoning (cf. symbolic) [22]. The nature of human information processing has recently been aptly described by Harari: "*[..] the mind is a flow of subjective experiences [..] made of interlinked sensations, emotions and thoughts [..]. When reflecting on it, we often try to sort the experiences into distinct categories such as sensations, emotions and thoughts [..]*". ([15], p. 123).

For humans and intelligent machines to jointly learn and perform a task, both should develop and maintain a common vocabulary of concepts and relations (challenge 1); to reason and communicate with each other (challenge 6), about the task and environment (challenge 3), their own perspective (challenge 4), and the perspective of the other (challenge 5). This means, for example, that an ePartner of a machine agent should be able to translate implicit sub-symbolic knowledge, that is acquired through associations, into symbolic concepts (see Fig. 3). Only then can this ePartner-agent communicate with other agents about it.

Perceptual, Cognitive, and Social Components

Machine agents and ePartner agents should be able to represent knowledge obtained from sensory experiences by building associative networks of perceptual inputs (challenge 3). This would allow the agent to, for example, perform image classification and object recognition.

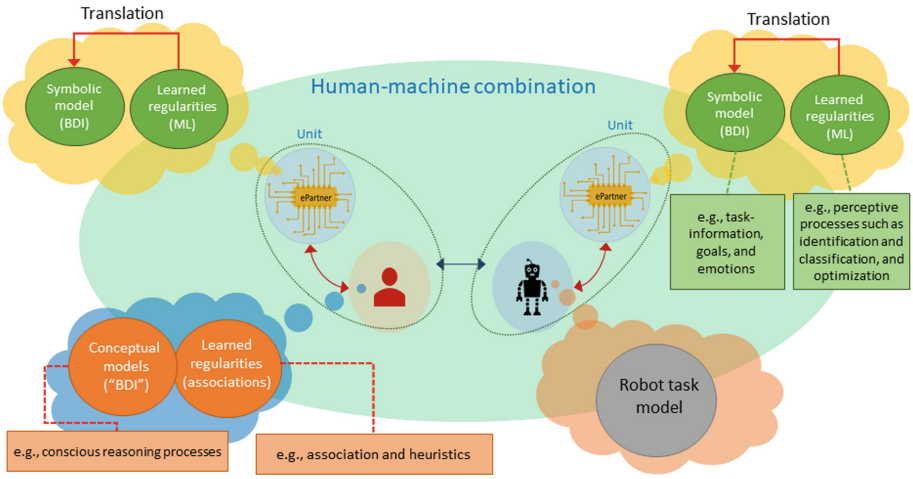


Fig. 3. A human-machine combination consisting of a human agent with its ePartner agent, and an intelligent robot (e.g., an UGV), also with its ePartner agent. All agents build and maintain their own mental models that contain learned regularities (acquired through e.g., Machine Learning), as well as symbolic knowledge (e.g., in terms of BDI). To enable communication, ePartner-agents should be able to translate sub-symbolic knowledge into symbolic terms. This symbolic model functions as a shared vocabulary for agents.

In the Human-AI co-learning concept, agents have different capabilities and they communicate in order to exploit their complementary strengths to the benefit of the team. Agents should therefore be able to show and share their status and intentions to their partners (challenge 6), demanding a mental model that allows them to express and explain their beliefs, goals, intentions, and actions in terms that are adequate for human understanding and appreciation (challenge 5) [41]. If necessary, the mental model can be expanded with computational models of emotion [23], enabling agents to take affective states into account when deciding upon which goals to pursue and which actions to perform. Some of the socially adaptive behavior of agents can be streamlined in advance by setting and agreeing upon work agreements. However, in order for agents to know when and how to adapt to changes, they should be able to continuously collect and update information about the individual team member(s) and the context (challenge 2). The agent's mental model should therefore have slots for social information; with strategies for obtaining information from the task context to fill and refine the value of these variables; and with algorithms to make social inferences from the data in the model (challenge 1).

3.3 Functional Challenges for Mental Models

Constructing mental models for agents that allow them to represent perceptual, cognitive-affective, and social knowledge, is only part of the challenge. Agents

should also have the capabilities to dynamically update and refine their models. This can be achieved in various ways, like internal consistency checks, deduction, induction, reasoning, and validation. The sections below discuss the challenges associated with establishing these functions.

Dynamic Mental Models

Human-AI co-learning demands mental models to be dynamic, because human and machine agents will generally not have a mutual understanding right from the start. Instead, understanding develops over time, from experiences and interactions during training and operations. Of course, there may be some prior experience in the form of memories, 'lessons learned', and assumptions in the human agent, as well as computational task models in the machine agent. Furthermore, the human may have provided the machine agent with personal data to make itself better known. But a deeper understanding and mutual awareness develops through prolonged collaboration, interaction, shared experiences, and feedback from the environment (see also Sect. 2.2). To facilitate these processes, the human should be instructed and trained for understanding an AI-agent, and the AI-agent should have the functionalities to develop an understanding of his human teammate (challenge 5).

Mental Models That Support Observability, Predictability, Explainability, and Directability

Humans are cognitively wired to automatically infer mental states from subtle behavioral cues expressed by other agents [17,34]. To enhance its observability [19] for a human partner, a machine agent should be able to express such information about its 'mental' state in a way that is easy to comprehend for its human partner (challenge 6). In addition, a machine agent should be able to infer its human partner's mental state from their behavior (challenge 5). Agents should also be able to use their theory of mind model to make predictions about the behavior of team partners. Comparisons with observed behavior should be used to validate the model, and to make adjustments if necessary.

Members of an effective human-human team try to detect and solve discrepancies in their mental models. They discover misunderstandings, diagnose the cause, and provide corrective explanations that gets the team back on track [12,37]). Likewise, machine agents too need to be explainable. They need to be able to generate explanations that shed light on the underlying causes of their actions, and are attuned to the characteristics of the receiving agent (challenge 6) [37]. Agents may form explanations reactively, in response to a request by a partner, but also pro-actively, when the agent anticipates that a partner may not understand its (choice of) behavior.

Directability refers to the property of agents to take over and delegate tasks, both reactively and pro-actively. The agent should be able to consult its model of the team (challenge 2), taking also into consideration the level of interpersonal trust (see Sect. 2.2).

4 Addressing the Challenges

In a successful hybrid team, humans and machines collaborate in an adaptive, dynamic, and personalized fashion. This requires that machine agents, just like humans, have mental models that contain information on the task context, their own role, and the role of others. Furthermore, human and machine agents should be able to express and clarify their mental states in a way that is easy to comprehend for their partner and that allows them to act in a coordinated and adaptive manner.

In this paper, we have proposed six challenges to achieve successful human-AI partnership in hybrid teams. These challenges should be addressed and tested in research. A good start would be, for example, to investigate how learning of an individual agent can be organized and supported, studying how this learning affects performance of others, first at the unit level and successively at the team level. Research questions would concern the construction, maintenance, and use of mental models (e.g., what information should an agent disclose to elicit adaptive responses from its partner or partners?, and what are the effects of different explanations by others on an agent's learning?).

In order to address the challenges, a suitable research simulation environment is needed. This needs to involve a task that is representative for a real world environment in which humans and intelligent technology jointly work together. Yet the research task should allow simple and unambiguous manipulation and control of demands on human-AI cooperation, and should allow the measurement of learning. A suitable research environment is needed that meets the following requirements: (1) control over what information is available. If some information about the task is unknown or uncertain to some agents, this requires them to communicate and to generate explanations that facilitate mutual understanding; (2) the opportunity to create hard interdependencies [19], compelling agents to cooperate because each have unique capabilities; (3) control over resources needed to carry out the task (e.g., imposing time limits); and (4) the opportunity to make task goals achievable in several ways. This feature requires agents of a unit or team to search for common ground on strategy, and to explore the division of roles and tasks that result in good collective performance. Earlier studies on human-AI collaboration have been using research environments that meet the requirements above, like Blocks World for Teams [20] and Hanabi [2].

We intend to use such environments to conduct experimental research into human-AI learning. As a first study we aim to investigate how a human and machine agent can evaluate their joint task performance in terms of lessons for the future (i.e., how to re-assign tasks to improve overall performance). Controlled studies in the lab are needed to design, implement and evaluate the principles for successful Human-AI Co-learning. In addition, these principles should be tested further in practical field settings (e.g., similar to experiments by De Greeff et al. [13] and Looije et al. [33]). For example, trust has been shown to be an important aspect of human-AI cooperation in real-life. The co-learning of humans and agents over a prolonged period of time may not only benefit performance, but also trust calibration [58].

Given the developments in society, the future will unequivocally demand humans and intelligent systems to work together. This paper addresses the challenges facing hybrid human-AI teams to acquire the strengths of human-human teams, and to exploit the unique benefits of intelligent technology at the same time.

Acknowledgments. This study has been funded by the Netherlands Ministry of Defence, under program V1801.

References

1. Bader, S., Hitzler, P.: Dimensions of neural-symbolic integration-a structured survey (2005). arXiv preprint cs/0511042
2. Bard, N., et al.: The Hanabi challenge: a new frontier for AI research (2019). arXiv preprint: arXiv:1902.00506
3. Bergstein, B.: AI isn't very smart yet. But we need to get moving to make sure automation works for more people (2017). https://www.technologyreview.com/s/609318/the-great-ai-paradox/
4. van den Bosch, K., Bronkhorst, A.: Human-AI cooperation to benefit military decision making. In: Proceedings of the NATO IST-160 Specialist' Meeting on Big Data and Artificial Intelligence for Military Decision Making, Bordeaux, France, 30 May–1 June 2018, S3-1/1-S3-1/12 (2018)
5. Bosse, T., Breebaart, L., Diggelen, J.V., Neerincx, M.A., Rosa, J., Smets, N.J.: Developing epartners for human-robot teams in space based on ontologies and formal abstraction hierarchies. Int. J. Agent-Oriented Softw. Eng. **5**(4), 366–398 (2017)
6. Bron, R., Endedijk, M.D., van Veelen, R., Veldkamp, B.P.: The joint influence of intra-and inter-team learning processes on team performance: a constructive or destructive combination? Vocations and learning, pp. 1–26 (2018)
7. Brooks, R.: The Seven Deadly Sins of AI Predictions (2017). https://www.technologyreview.com/s/609048/the-seven-deadly-sins-of-ai-predictions/
8. Converse, S., Cannon-Bowers, J., Salas, E.: Shared mental models in expert team decision making. In: Individual and Group Decision Making: Current Issues Issues, p. 221 (1993)
9. Costa, A.C., Fulmer, C.A., Anderson, N.R.: Trust in work teams: an integrative review, multilevel model, and future directions. J. Organ. Behav. **39**(2), 169–184 (2018)
10. Evans, J.S.B.: Dual-processing accounts of reasoning, judgment, and social cognition. Annu. Rev. Psychol. **59**, 255–278 (2008)
11. Fryer, L.K., Nakao, K., Thompson, A.: Chatbot learning partners: connecting learning experiences, interest and competence. Comput. Hum. Behav. **93**, 279–289 (2019)
12. de Graaf, M., Malle, B.F.: How people explain action (and autonomous intelligent systems should too). In: AAAI Fall Symposium on Artificial Intelligence for Human-Robot Interaction (2017)
13. de Greeff, J., Hindriks, K., Neerincx, M.A., Kruijff-Korbayova, I.: Human-robot teamwork in USAR environments: the TRADR project. In: Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts, pp. 151–152. ACM (2015)

14. de Greeff, J., Mioch, T., van Vught, W., Hindriks, K., Neerincx, M.A., Kruijff-Korbayová, I.: Persistent robot-assisted disaster response. In: Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, pp. 99–100. ACM (2018)
15. Harari, Y.N.: Homo Deus: A Brief History of Tomorrow. Random House (2016)
16. Harbers, M., Neerincx, M.A.: Value sensitive design of a virtual assistant for workload harmonization in teams. Cogn. Technol. Work **19**(2–3), 329–343 (2017)
17. Heider, F.: The Psychology of Interpersonal Relations. Psychology Press, New York (1958)
18. Janis, I.L.: Groupthink. IEEE Eng. Manag. Rev. **36**(1), 36 (2008)
19. Johnson, M., et al.: Coactive design: designing support for interdependence in joint activity. J. Hum. Robot Interact. **3**(1), 43–69 (2014)
20. Johnson, M., Jonker, C., van Riemsdijk, B., Feltovich, P.J., Bradshaw, J.M.: Joint activity testbed: blocks world for teams (BW4T). In: Aldewereld, H., Dignum, V., Picard, G. (eds.) ESAW 2009. LNCS (LNAI), vol. 5881, pp. 254–256. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-10203-5_26
21. Johnson-Laird, P.N.: Mental models in cognitive science. Cogn. Sci. **4**(1), 71–115 (1980)
22. Kahneman, D., Egan, P.: Thinking, Fast and Slow, vol. 1. Farrar, Straus and Giroux, New York (2011)
23. Kaptein, F., Broekens, J., Hindriks, K.V., Neerincx, M.: CAAF: a cognitive affective agent programming framework. In: Traum, D., Swartout, W., Khooshabeh, P., Kopp, S., Scherer, S., Leuski, A. (eds.) IVA 2016. LNCS (LNAI), vol. 10011, pp. 317–330. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-47665-0_28
24. Kenny, P., et al.: Building interactive virtual humans for training environments. In: Proceedings of I/ITSEC, vol. 174, pp. 911–916 (2007)
25. Kim, Y., Baylor, A.L.: based design of pedagogical agent roles: a review, progress, and recommendations. Int. J. Artif. Intell. Educ. **26**(1), 160–169 (2016)
26. Klein, G., Woods, D.D., Bradshaw, J.M., Hoffman, R.R., Feltovich, P.J.: Ten challenges for making automation a "team player" in joint human-agent activity. IEEE Intell. Syst. **19**(6), 91–95 (2004)
27. Knight, W.: More evidence that humans and machines are better when they team up - MIT Technology Review.pdf (2017). https://www.technologyreview.com/s/609331/more-evidence-that-humans-and-machines-are-better-when-they-team-up/
28. Kruijff-Korbayová, I., et al.: TRADR project: long-term human-robot teaming for robot assisted disaster response. KI-Künstliche Intell. **29**(2), 193–201 (2015)
29. Lake, B.M., Ullman, T.D., Tenenbaum, J.B., Gershman, S.J.: Building machines that learn and think like people. Behav. Brain Sci. **40**, e253 (2017)
30. Lemaignan, S., Warnier, M., Sisbot, E.A., Clodic, A., Alami, R.: Artificial cognition for social human-robot interaction: an implementation. Artif. Intell. **247**, 45–69 (2017)
31. Lin, J., Spraragen, M., Zyda, M.: Computational models of emotion and cognition. In: Advances in Cognitive Systems. Citeseer (2012)
32. Liszka-Hackzell, J.J.: Prediction of blood glucose levels in diabetic patients using a hybrid AI technique. Comput. Biomed. Res. **32**(2), 132–144 (1999)
33. Looije, R., Neerincx, M.A., Cnossen, F.: Persuasive robotic assistant for health self-management of older adults: design and evaluation of social behaviors. Int. J. Hum. Comput. Stud. **68**(6), 386–397 (2010)
34. Malle, B.F.: How the Mind Explains Behavior. Folk Explanation, Meaning and Social Interaction. MIT Press, Cambridge (2004)

35. Mathieu, J.E., Heffner, T.S., Goodwin, G.F., Salas, E., Cannon-Bowers, J.A.: The influence of shared mental models on team process and performance. J. Appl. Psychol. **85**(2), 273 (2000)
36. Mayer, R.C., Davis, J.H., Schoorman, F.D.: An integrative model of organizational trust. Acad. Manag. Rev. **20**(3), 709–734 (1995)
37. Miller, T.: Explanation in artificial intelligence: insights from the social sciences. In: Artificial Intelligence (2018)
38. Mioch, T., Peeters, M.M., Nccrincx, M.A.: Improving adaptive human-robot cooperation through work agreements. In: 2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), pp. 1105–1110. IEEE (2018)
39. Morrison, J.E., Meliza, L.L.: Foundations of the after action review process. Technical report, Institute for Defense Analyses, Alexandria, VA (1999)
40. Neerincx, M., et al.: The mission execution crew assistant: improving human-machine team resilience for long duration missions. In: Proceedings of the 59th International Astronautical Congress (IAC 2008) (2008)
41. Neerincx, M.A., van der Waa, J., Kaptein, F., van Diggelen, J.: Using perceptual and cognitive explanations for enhanced human-agent team performance. In: Harris, D. (ed.) EPCE 2018. LNCS (LNAI), vol. 10906, pp. 204–214. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91122-9_18
42. Nikolaidis, S., Hsu, D., Srinivasa, S.: Human-robot mutual adaptation in collaborative tasks: models and experiments. Int. J. Robot. Res. **36**(5–7), 618–634 (2017)
43. Parasuraman, R., Barnes, M., Cosenzo, K., Mulgund, S.: Adaptive automation for human-robot teaming in future command and control systems. Technical report, Army Research Lab Aberdeen proving ground MD Human Research and Engineering Directorate (2007)
44. Patterson, R.E., Pierce, B.J., Bell, H.H., Klein, G.: Implicit learning, tacit knowledge, expertise development, and naturalistic decision making. J. Cogn. Eng. Decis. Mak. **4**(4), 289–303 (2010)
45. Premack, D., Woodruff, G.: Does the Chimpanzee have a theory of mind? Behav. Brain Sci. **1**(4), 515–526 (1978)
46. Reber, A.S.: Implicit learning and tacit knowledge. J. Exp. Psychol. Gen. **118**(3), 219 (1989)
47. Salas, E.: Team Training Essentials: A Research-Based Guide. Routledge, London (2015)
48. Salas, E., Reyes, D.L., McDaniel, S.H.: The science of teamwork: progress, reflections, and the road ahead. Am. Psychol. **73**(4), 593 (2018)
49. Siegel, A.W., Schraagen, J.M.: Team reflection makes resilience-related knowledge explicit through collaborative sensemaking: observation study at a rail post. Cogn. Technol. Work **19**(1), 127–142 (2017)
50. Silver, D., et al.: Mastering the game of go without human knowledge. Nature **550**(7676), 354 (2017)
51. Stout, R.J., Salas, E., Carson, R.: Individual task proficiency and team process behavior: what's important for team functioning? Mil. Psychol. **6**(3), 177–192 (1994)
52. Stout, R.J., Cannon-Bowers, J.A., Salas, E.: The role of shared mental models in developing team situational awareness: implications for training. In: Situational Awareness, pp. 287–318. Routledge (2017)

53. Teo, G., Wohleber, R., Lin, J., Reinerman-Jones, L.: The relevance of theory to human-robot teaming research and development. In: Savage-Knepshield, P., Chen, J. (eds.) Advances in Human Factors in Robots and Unmanned Systems. AISC, vol. 499, pp. 175–185. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-41959-6_15
54. Tielman, M.L., Neerincx, M.A., Bidarra, R., Kybartas, B., Brinkman, W.P.: A therapy system for post-traumatic stress disorder using a virtual agent and virtual storytelling to reconstruct traumatic memories. J. Med. Syst. **41**(8), 125 (2017)
55. Tielman, M.L., Neerincx, M.A., van Meggelen, M., Franken, I., Brinkman, W.P.: How should a virtual agent present psychoeducation? Influence of verbal and textual presentation on adherence. Technol. Health Care **25**, 1–16 (2017). Preprint
56. Tsaih, R., Hsu, Y., Lai, C.C.: Forecasting s&p 500 stock index futures with a hybrid ai system. Decis. Support Syst. **23**(2), 161–174 (1998)
57. Van Diggelen, J., Neerincx, M., Peeters, M., Schraagen, J.M.: Developing effective and resilient human-agent teamwork using team design patterns. IEEE Intell. Syst. **34**(2), 15–24 (2018)
58. de Visser, E.J., et al.: Longitudinal trust development in human-robot teams: models, methods and a research agenda. IEEE Trans. Hum. Mach. Syst., 1–20 (2018)
59. Werkhoven, P., Kester, L., Neerincx, M.: Telling autonomous systems what to do. In: Proceedings of the 36th European Conference on Cognitive Ergonomics, p. 2. ACM (2018)
60. Wiltshire, T.J., Fiore, S.M.: Social cognitive and affective neuroscience in human-machine systems: a roadmap for improving training, human-robot interaction, and team performance. IEEE Trans. Hum. Mach. Syst. **44**(6), 779–787 (2014)
61. Wooldridge, M., Jennings, N.R.: Agent theories, architectures, and languages: a survey. In: Wooldridge, M.J., Jennings, N.R. (eds.) ATAL 1994. LNCS, vol. 890, pp. 1–39. Springer, Heidelberg (1995). https://doi.org/10.1007/3-540-58855-8_1
62. Xiao, Z., et al.: A deep learning-based segmentation method for brain tumor in MR images. In: 2016 IEEE 6th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS), pp. 1–6. IEEE (2016)