Engineering 25 (2023) 77–100

Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Artificial Intelligence—Review

Communicative Learning: A Unified Learning Formalism

Luyao Yuan [a,b,*], Song-Chun Zhu [a,c,d,*]

[a] Beijing Institute for General Artificial Intelligence, Beijing 100086, China
[b] Department of Computer Science, University of California, Los Angeles, CA 90024, USA
[c] Department of Automation, Tsinghua University, Beijing 100084, China
[d] Institute for Artificial Intelligence, Peking University, Beijing 100871, China

ARTICLE INFO

Article history:
Received 4 February 2022
Revised 5 September 2022
Accepted 27 October 2022
Available online 31 March 2023

Keywords:
Artificial intelligence
Cooperative communication
Machine learning
Pedagogy
Theory of mind

ABSTRACT

In this article, we propose a communicative learning (CL) formalism that unifies existing machine learning paradigms, such as passive learning, active learning, algorithmic teaching, and so forth, and facilitates the development of new learning methods. Arising from human cooperative communication, this formalism poses learning as a communicative process and combines pedagogy with the burgeoning field of machine learning. The pedagogical insight facilitates the adoption of alternative information sources in machine learning besides randomly sampled data, such as intentional messages given by a helpful teacher. More specifically, in CL, a teacher and a student exchange information with each other collaboratively to transmit and acquire certain knowledge. Each agent has a mind, which includes the agent's knowledge, utility, and mental dynamics. To establish effective communication, each agent also needs an estimation of its partner's mind. We define expressive mental representations and learning formulation sufficient for such recursive modeling, which endows CL with human-comparable learning efficiency. We demonstrate the application of CL to several prototypical collaboration tasks and illustrate that this formalism allows learning protocols to go beyond Shannon's communication limit. Finally, we present our contribution to the foundations of learning by putting forth hierarchies in learning and defining the halting problem of learning.

1. Introduction

Better than a thousand days of diligent study is one day with a great teacher.
听君一席话胜读十年书。 —Chinese proverb
When I walk along with two others, they may serve me as my teachers.
三人行, 必有我师焉。 —Confucius

1.1. Objective: A unifying formalism

The recent surge of statistical and machine learning has enabled artificial intelligence (AI) to achieve impressive performance in certain specific and well-defined tasks. However, current machine learning paradigms also demonstrate several shortcomings: a demand for large quantities of training data, uninterpretable and noncommunicable representations, and deficient generality to novel tasks and unknown situations. These machine learning methods belong to a "big data for small tasks" paradigm [1] drastically different from human learning, which communicates a wide range of daily tasks effectively using small data, namely, "small data for big tasks." Human learning, taking place in pedagogical environments, also entails many layers of cognitive infrastructures and diverse protocols between multiple participants. Yet, such complexity and sophistication are usually simplified in common machine learning methods. To fill in the gap between prevailing machine learning approaches and human learning, in this article, we propose the concept of communicative learning (CL). As a unifying formalism, CL involves multiple agents: A teacher and a student communicate with each other in the process of teaching and learning. In CL, an agent's mental representation includes a set of minds:

- An egocentric mind that consists of the agent's current belief of the knowledge of interest, its utility, and its dynamic functions;
- A mind that estimates its partner's egocentric mind;

* Corresponding authors.
E-mail addresses: comm.learn.ml@gmail.com (L. Yuan), s.c.zhu@pku.edu.cn (S.-C. Zhu).

https://doi.org/10.1016/j.eng.2022.10.017

- A common mind holding the common ground established within the group;
- A "God's mind" bearing the actual fact of the world.

These mental components then jointly drive learning and communication. With CL, we pose learning as a communication process and demonstrate its advantage over non-CL methods. Furthermore, we show that this learning formalism encompasses and goes beyond existing learning methods. We use the perspective of mutual reasoning between a teacher and student to survey prior works on various kinds of machine learning algorithms. We illustrate that, despite their diversity, many types of teachers and students proposed thus far can be characterized as special cases of CL. This offers a unifying lens for the integration of cooperative pedagogy [2] and machine learning, helping readers to better understand and compare prior methods.

1.2. Cognitive infrastructure for CL

Communicative behaviors are so prevailing in human societies that most of us take them for granted, without realizing the complexity of the cognitive infrastructure that permits even the simplest communications to occur. Even intentional signals, which are considered to be a rudimentary form of human communication, are extremely rare in the biological world, perhaps confined to primates or even great apes (see Chapter 2 in Ref. [3]). Far beyond simple signaling or deliberately informing others, human communication is a complicated system that aims to establish joint attention and common ground for the completion of shared goals; it is motivated by cooperation norms among people and enabled by communicative conventions and cognitive infrastructure supporting recursive cooperative reasoning (see Chapter 3 in Ref. [3]). Human learning, as a lifelong cognitive process of communicating with the physical and social world, also operates in such a cooperative framework. Its sophistication, effectiveness, and complexity give rise to human intelligence—a phenomenon that AI is inspired to replicate.

Decades of studies in cognitive psychology [4,5] and anthropology and communications [3] have revealed that human communication and learning are built on many layers of cognitive infrastructure and protocols. To account for their complexity and sophistication, we formulate the mental representation of CL and show the key components between two agents in Fig. 1. Both of the agents can be human or machine, teacher or student in an equal and symmetric setting—that is, they can exchange roles by turns. This representation has the following properties:

(1) Theory of mind (ToM)[†] [6] representations, which engage six minds between the teacher and the student:

- G: the oracle in "God's mind;" this mind holds the actual state of the world and evolves according to the world transition model.
- P_t: the mind of the teacher A; this mind holds what the teacher, A, knows about the world, including the state, the model, A's utility and policy, and so forth.
- Q_t: the mind of the student B; this mind holds what the student, B, knows about the world, and is a counterpart of P_t.
- $\widehat{Q}_t$: what A thinks B knows; this mind helps the teacher to conduct cooperative pedagogy.
- $\widehat{P}_t$: what B thinks A knows; this mind helps the student to better capture the teacher's instructions.
- C_t: the common mind that both A and B know, and know that each other knows, and know that each other knows they know, and so forth (see Chapter 4 in Ref. [7]).

[†] The ability to attribute mental states such as beliefs, intents, desires, emotions, and knowledge to others.

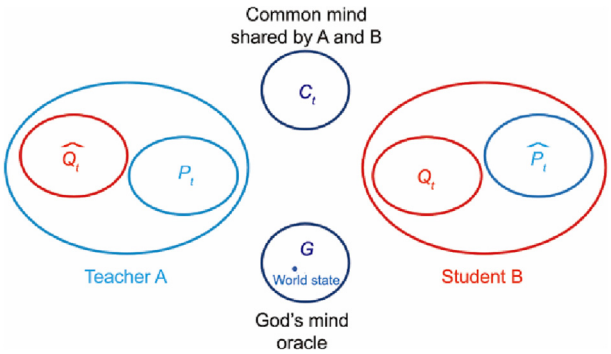

Fig. 1. Key representations of CL include six minds that evolve over t time steps or messages. G: the oracle in "God's mind;" P_t: the mind of the teacher A; Q_t: the mind of the student B; $\widehat{P}_t$: what B thinks A knows; $\widehat{Q}_t$: what A thinks B knows; C_t: the common mind that both A and B know, and know that each other knows, and know that each other knows they know, and so forth.

The discrepancy between the six minds drives the communication and learning process through deliberated messages instead of randomly sampled examples.

(2) An adaptive common mind and learning protocol: The common mind, C_t is the basis for future communication and learning. The more they share in the common mind, C_t, the more effectively the two agents communicate. For example, they can teach new concepts based on analogy [4] and modifications of existing concepts in their common knowledge. Suppose that both agents know what a dog looks like; then A can teach the concept of wolves to B by saying that a wolf is like a dog (i.e., copying its graphical structure and attributes) except for some features of appearance and behavior. Thus, CL allows the common knowledge to grow and integrates it into the learning protocol.

Our insight is that, armed with appropriate infrastructure sufficient for these properties, CL makes the following contributions to the field of machine learning:

- It provides a unifying framework for learning, which embraces existing machine learning methods as its special cases in several axes—including supervised versus unsupervised, passive versus active, and observational versus causal experimentation—and eases the development of new learning protocols under novel conditions.
- It integrates pedagogy into machine learning, which facilitates the learning through a helpful teacher and cooperative communication.
- It generalizes existing communication theories, such as Shannon's communication limit [8], Valiant and Vapnik's statistical learning framework [9], and so forth, to a machine learning process and proposes more efficient learning protocols.
- It studies the fundamental limits and the halting problem of learning, which determine the stopping conditions of a learning process at various equilibria among different minds.

To summarize, the CL representations enable the teacher to choose messages for the student according to various criteria [10–12], given her intended teaching content, and the student can update his belief according to his estimation of the teacher's message selection mechanism [13–15]. With distinctive teaching criteria, we can unify diverse teachers, from the passive oracle to the cooperative pedagogue. The latter delivers significantly more efficient learning protocols than those described by conventional communication and learning theories, whose fundamental assumption is random sampled data. In Section 3, we elaborate the CL representation and rigorously define the components of each mind.

1.3. A unifying framework for machine learning

CL is proposed as a unifying framework, where existing machine learning algorithms can be shown as its special cases. Dating back to the 1960s, machine learning was put forth to enable computers to recognize and capture patterns from data [16–18]. These days, with the availability of extensive data and computational power, machine learning algorithms are fulfilling their original goal of data understanding better and better. Complicated models can be exploited for various tasks, from image classification [19,20], object detection [21–23], and sentence generation [24] to exceeding human-level game playing [25,26], and so on.

However, most of the prevailing machine learning methods focus on the optimization of individual learners, relying purely on unilateral experiences—either passive interaction history from a Markov decision process (MDP) [25,26], labeled training examples from a data distribution [19,27], answers to active queries provided by an oracle [28,29], or demonstrations from an expert [30]. Only recently has the advantage of pedagogical teachers over randomly sampled data or optimal task completion trajectories from experts been shown in Bayesian concept learning [13,31–35] and in learning from demonstration (LfD) [15,36,37]. Meanwhile, machine teaching algorithms are starting to model cooperative teachers giving instructions in a continuous parameter space and large datasets to enhance learning efficiency [38–42] or robustify the model against noisy labels [43,44], albeit only with simple learner models. In Fig. 2, we compare some typical learning paradigms.

1.4. Integrating pedagogical reasoning

Another merit of CL is that it enables the integration of human pedagogy into machine learning. When one pictures human learning—whether the learning of a toddler at home or a student at school—the scenario often involves two parties: a teacher and a student. The teacher tries to impart her knowledge to the student using the most helpful teaching material, while the student actively absorbs the information to achieve the learning goal as efficiently as possible. Taking place in a multiagent system, this type of learning is a communicative process, which is referred to as pedagogy [2]. In machine learning, however, the role of the helpful teacher is usually replaced by training data (in the form of a dataset or interaction experience) sampled from some random process [18]. Suppose that we draw an analogy: Without a cooperative teacher, machine learning algorithms act like scientists that endeavor to detect, explain, and utilize the pattern of the world from their observations of randomly occurring phenomena, such as apples dropping from trees, stars shining, and rain falling. The most common type of learning in human society is not scientific discovery but cooperative pedagogy, which is invoked across language, cognitive development, and cultural anthropology to explain people's ability to effectively transmit information and accumulate knowledge [42,45].

There has been sufficient experimental evidence in cognitive science to justify people's ability and tendency to teach and learn differently in pedagogical situations. From infants' and toddlers' ability to distinguish different sampling processes [46] and their awareness of pedagogical behavior [47–49] to the application of pedagogical inference in word learning for both children and adults [50], these studies illustrate that human learners are sensitive to the distinction between pedagogical teaching material and random examples. From a very young age, humans can capture additional information when they are in pedagogical settings [31]. Furthermore, besides being capable learners, children can also learn how to be a good teacher and generate helpful evidence for others when trying to reveal the mechanism of toys [51].

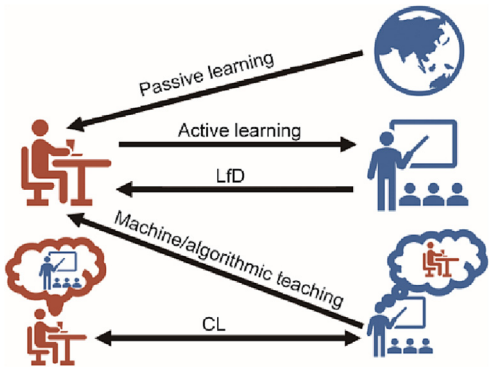

Fig. 2. A comparison of different learning/teaching protocols. Blue characters represent teachers and red characters are students. The bubble refers to modeling one's partner.

The main difference between a scientist analyzing data and a student in a pedagogical situation is the existence of a cooperative teacher, who can maximize communication efficiency by interacting adaptively with the student. More specifically, the teacher adjusts her teaching method for different students, and the student—after becoming familiar with the teacher's instruction mechanism—can infer the teacher's intention and learn faster [10,18,48,52,53]. Recently, realizing the conundrum of data hunger and "big data for small tasks" [1] for conventional machine learning algorithms (especially compared with the efficiency of human learning, which can be quickly consummated with very limited examples [31,48,54]), researchers have begun to integrate pedagogy into machine learning algorithms [13,15,31–42]. Nonetheless, compared with human pedagogy, these works lack a sophisticated student model that can accommodate the teacher's cooperation into his learning and whose learning differs from learning from passive data. Recursive cooperative inference models are proposed in Refs. [45,55], with both the teacher and the student having ToM [6]. Yet, the analysis of these works is confined to Bayesian concept learning with a finite and relatively small dataset and hypothesis space. In Section 3.3, we illustrate how CL facilitates the combination of human pedagogy and machine learning, as well as the development of more advanced learning algorithms. The remainder of this paper is structured as follows: In Section 2, we clarify the relationship between learning and communication and motivate a CL paradigm. In Section 3, we introduce the CL formalism, starting with agent modeling and agents' necessary mental representations, which are followed by mathematical definitions of the learning dynamics. We also show how existing learning algorithms are special cases of CL and their CL grounding in Section 3.4 (with more details in Sections S1 and S3 in Appendix A). Then, we demonstrate the applicability and necessity of the CL formalism with a few concrete examples in Section 4, including a referential game and a real-time human–robot interaction (HRI) task example. Afterward, Section 5 illustrates the theoretical contributions of the CL formalism. We put forth a new representation of learning and give a learning protocol beyond the Shannon communication limit in Section 5.1. In Section 5.2, we tease out the three hierarchies of machine learning and define the fundamental halting problem of learning. Finally, we conclude by describing the implications of CL.

2. Inspect learning in a communication framework

Let us return to the scientists versus students analogy. It does not take much argument to convince people that learning is a communication process, because most of us were once (or still are) students and experienced learning in academic settings, by which we

acquired knowledge, skills, and values through communicative interactions with teachers. Within Western culture, this form of learning is said to originate from Socrates, who exposed the idea of using dialogues between individuals to elicit and impart wisdom and to transfer knowledge from the teacher's mind to the student's. In fact, the one-way communication from nature to a scientist—if we are willing to assume the existence of an omniscient being such as God—can also be viewed as transferring knowledge from God's mind to the scientist's mind; it is just that the messages from God are not as decipherable as those from Socrates (at least to most people). More generally, we can thus broadly interpret learning as having information delivered from one mind to another. Interestingly, a textbook definition of communication is exactly that: the act of delivering. Communication and learning are clearly connected, but what is the relationship between communication and learning? An obvious way to answer this question is to look at information theory [8] and statistical learning theory [56]—two well-developed disciplines that formally study communication and learning from a mathematical viewpoint. As we will see, they have a great deal in common.

2.1. The connection between communication and learning

Information theory was established by Claude Shannon in 1948 [8] as a framework to account for communication over, say, a telephone wire. As shown in Fig. 3, in this framework, the sender and receiver share a codebook, and the messages refer to some world state w, such as a parse graph of an indoor scene or a semantic of a paragraph. Such a communication system is for reproducing—either exactly or approximately—a message selected at other points. Loosely speaking, it involves an information source that selects a message from a set of possible messages and then encodes it into a sequence of communication symbols suitable for transmission. This sequence is subsequently sent over a possibly noisy medium to a receiver that decodes and reconstructs the original message according to a predefined coding scheme. Significantly, Shannon showed that, depending on the characteristics of the transmission medium, there is an upper limit on the rate (e.g., bits per second) at which messages can be reliably transmitted through such a system. This limit is termed the channel capacity. Every time a message is received from the channel, the receiver's uncertainty about the possible worlds will decrease by an amount bounded by the channel capacity. Suppose that we denote the set of possible worlds at time t as $\mathscr{W}^t$. Then, the information gain, IG^{t+1}, brought by a message at time $t+1$ can be represented as follows:

$$\mathrm{IG}^{t+1} = \log_2 \frac{1}{\left|\mathscr{W}^{t+1}\right|} - \log_2 \frac{1}{\left|\mathscr{W}^{t}\right|} = \log_2 \frac{\left|\mathscr{W}^{t}\right|}{\left|\mathscr{W}^{t+1}\right|} \tag{1}$$

One drawback of Shannon's theory is that it intentionally leaves out the "semantics" or "meanings" of messages. The sender and receiver are assumed to share common ground to make sense of the messages outside of this framework. The limits of coding efficiency and channel capacity are based on a protocol that does not model the mental states and motives of the agents sending messages.

2.2. Statistical machine learning

Compared with information theory, statistical learning theory—although it is relatively young—goes one step further. As per the most well-known probably approximately correct (PAC) model proposed by Valiant [9], learning seeks to accurately acquire a concept selected randomly by looking at some training data. Fig. 4 shows the setting of PAC learning. Formally, the learner tries to

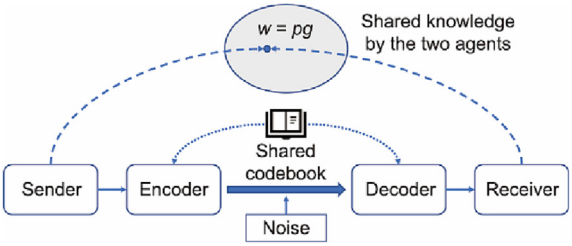

Fig. 3. Shannon's diagram of a communication system. The shared codebook fulfills the purpose of a common ground and acts as the norm of communication in the cooperative communication model. w: world state, represented by parse graph (pg).

learn a concept c, which is defined as a set in a state space Ω [9] or a probability model θ in a hypothesis space Θ ($\theta \in \Theta$) [57], using M samples, $\{(I_i, c_i), i = 1, \ldots, M\}$ that are drawn randomly from an external world. Here (I_i, c_i) are observations of certain form, such as images, labeled class c_i. Learning is driven by a predefined utility or loss function u. The PAC learning theory bounds the number of training data, $n(\epsilon, \delta)$, needed to learn the concept with error $\leq \epsilon$ and confidence $> 1 - \delta$.

Therefore, by equating a concept with a message and training data with communication symbols, we find a strong parallel between Shannon's communication model and PAC learning. To continue the previous metaphor, learning describes the communication of concepts from one mind to another. More specifically,

- Acting like an information source, nature selects a concept from a class according to a certain probability. It then uses a sequence of instances to encode the concept with binary labels to form consistent training examples.
- Like signals transported over a transmission medium, labels may then be subject to noise and bit-flip.
- Upon seeing possibly noisy training examples, a machine runs a learning algorithm to decode and reconstruct the target concept [58].
- Analogous to channel capacity, in learning, there is an upper bound on how quickly concepts can be transmitted with training examples. This is measured by sample complexity—that is, the minimum number of training examples a machine needs to see before it can reliably identify any intended concept.

In summary, statistical learning resembles a communication framework that emphasizes decoding. Hence, when modeling a

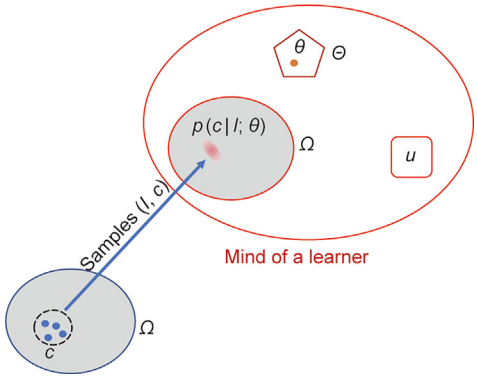

Fig. 4. Diagram for statistical learning theory. The learner passively receives the examples from the physical world. There are only egocentric mental components in the learner's mind. Here Ω represents the state space and c is a concept defined as a set in a state space. (I, c) is a random sample complying with the concept. Given the received samples, the learner can form a belief, p, of the concept. The hypothesis $\in \Theta$, in the hypothesis space Θ, determines the belief and the predefined utility function, u, drives the learning process.

learning process, it shares the same drawback as Shannon's communication channel. Can we go beyond Shannon's communication limit in machine learning if we are to overcome the lack of mental modeling? The CL formalism sheds light on this question, where messages are selected after deliberations and reflection using ToM representations and carry extra information that is recoverable in a more effective communication protocol; that is, agents are capable of "reading between the lines."

2.3. Going beyond Shannon's and Valiant's frameworks

Now that we have reviewed the connection between foundations of information theory and statistical learning especially PAC learning, keen readers must have realized that Valiant's learning model, despite its theoretical appeal, has rather limited explanatory power outside of the scientist's learning scenario. This model makes an inefficient assumption that examples in learning are random samples, whereas—in most real-life cases—learning is a communication process in which examples are deliberate messages uttered that reflect the mental states of the student and the teacher. In addition, the lack of pedagogy at the teacher's end eliminates the possibility of having a student that is anything other than a completely passive learner, whose model is much more rudimentary than our learning experience as human beings. Since we have agency and are guided by purposes, we not only engage in the interpretation of received information but also actively ask questions to dispel doubts. Furthermore, assuming that other people are similarly endowed and motivated, we often question why others behave in a certain way and take into account their possible reactions to our moves before we act. These crucial elements are missing from Valiant's model.

Here, we can show a toy example to illustrate the advantage of deliberate messages over random sampling. Suppose that there are four sets of numbers, $\{1,2,3\}$, $\{4,5,6\}$, $\{4,5\}$, and $\{1,2,4,6\}$, and the teacher can inform the student which one of the four is the target set by sending numbers that belong to the target set. Let us consider the case when the third set $\{4,5\}$ is the target. For a student learning from random sampling, it takes many repetitive numbers to shape his belief and eliminate the second set, which happens to be the superset of the target. The learner is only able to probabilistically eliminate the second set after constantly not receiving the number 6.

However, if the student is learning from a cooperative teacher, more efficient protocols can be established between the agents; in fact, some can even accomplish the deterministic identification of the target with only one message. For example, a cooperative teacher will use the number 3 to uniquely identify the first set, because no other sets include 3. Thus, the numbers 1 and 2 become available to identify the fourth set, since the teacher would choose 3 if the target was the first set. Similarly, the number 6 is freed from the fourth set and can be used to indicate the second set, leaving 4 and 5 as unique identifiers of the third set (see Fig. 5 [59] for a comparison between statistical learning from random data and this pedagogical learning). Such a learning protocol is only possible between cooperative and pedagogical agents.

Therefore, a model for Socratic learning via dialogue goes beyond just PAC learning—that is, the one-way communication of concepts: It involves meaning and intention. This observation is not new. As Shannon himself acknowledged [8], "the messages frequently have meaning; that is, they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem." In a follow-up work [60], Weaver also pointed out that a broader understanding of communication should include all the procedures by which one mind may affect

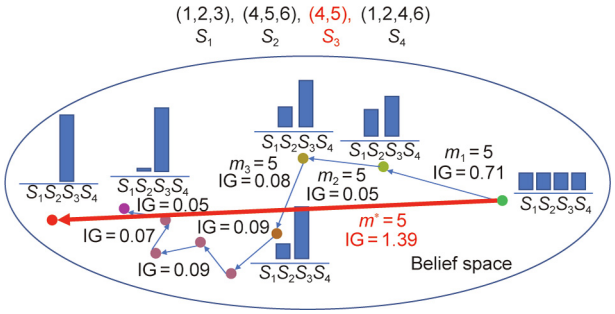

Fig. 5. The student's belief-update processes when learning from random sampling (blue trajectory) and from a cooperative teacher (red trajectory). It is possible to transmit the target set $\{4,5\}$ with a single message, because the recursive teaching dimension (RTD) [59] of this concept space is the number 1, which lower bounds the regular teaching dimension (TD). Concretely, the student knows that the teacher will send the number 3 for S_1; thus, the number 1 or 2 indicates S_4. Similarly, the number 6 can only mean S_2, making 5 a unique identifier for S_3. The prerequisite of such a recursive protocol is that the teacher is pedagogical, and that the student is aware of her helpfulness. Here m_1, m_2, and m_3 denotes the first, second, and third random message. m^* represents the message from the cooperative teacher. More concept classes with different TDs can be found in Table S1 in Appendix A.

another. More specifically, he suggested that general communication problems be considered at three levels:

(1) How accurately can the symbols of communication be transmitted? (the technical problem);

(2) How precisely do the transmitted symbols convey the desired meaning? (the semantic problem);

(3) How effectively does the received meaning affect conduct in the desired way? (the effectiveness problem).

Although it lies at the lowest level, the technical problem of transmitting symbols—or concepts for that matter—leads to insight in that it helps identify several key ingredients to enable successful communication, at all levels. However, symbol transmission only provides the necessary postal infrastructure for participants of a general communication process to convey meaning and express desire. Indeed, human communication is far more complex. As an illustration of how meaning construction (level 2) can be decoupled from transmitted symbols (level 1), it suffices to imagine the different reactions to the same television images of a football match by the fans of opposing sides. Furthermore, even if there is only one possible interpretation of a particular message, in human communication, the receiver need not simply accept but may alternatively ignore or oppose a message (level 3). The heated political discourses in the United States on climate change are a prime example of this phenomenon.

Since our goal is to model learning through communication in the broad sense, in light of the prior work, we must account for the two higher levels—and primarily the semantic problem level—of communication. To understand how level-2 communication (in semantics) is conducted among human beings, we refer to an account by Michael Tomasello from an evolutionary and cognitive perspective [3]. According to Tomasello, the key to the success of human communication is a common ground shared by its participants, which includes joint attention, shared experience, and common cultural knowledge. A common ground provides the critical situational, social, political, cultural, and historical context for people to construct meaning based on their received communication symbols.

To illustrate this point, consider an example given in Ref. [3]. Imagine that you and I are walking toward the library and, out of the blue, I point in the direction of some bicycles leaning against the library wall. Your reaction will very likely be "Huh?", because it is difficult for you to know which aspect of the situation I am

indicating or why I am doing so, since pointing on its own means nothing. However, if you just broke up with your boyfriend in a particularly nasty way, we both know this mutually, and one of the bicycles is his, which we also both know mutually, then the same pointing gesture in the same physical situation might mean something very complex, such as, "Your boyfriend's already at the library (so perhaps we should skip it)." On the other hand, if one of the bicycles is the one that we both know mutually was stolen from you recently, then the exact same pointing gesture will mean something completely different. Or perhaps we have been wondering together whether the library is open at this late hour; then, indicating the presence of many bicycles outside can be a sign that the library is open.

It is important to point out that the common ground is not a new notion, because it already existed in Shannon's communication model and Valiant's PAC learning model. In traditional communication, the common ground is represented by the common codebook shared by the sender and receiver, whereas in PAC learning, the common ground is the common class of concepts. However, what appears to be novel in human communication is that the common ground is jointly created and selected by the participants, rather than by a third person that sets up a communication system or a learning algorithm.

These observations highlight the connection between learning and communication. The resemblance between machine learning algorithms and the scientist's way of learning—that is, one-way communication—identifies the need for cooperative pedagogy and motivates our CL formalism. In the next section, we will show the mental representation of the teacher and the student, who communicate with each other to achieve efficient pedagogy.

3. Communicative learning

In this section, we define the CL formalism. We start with the infrastructure of CL and the mental representation of the teacher and the student. As discussed in Section 1.2, the infrastructure must have sufficient expressiveness to accommodate the common mind and ToM. Then, we present the dynamics of learning in Section 3.3 to illustrate how the learning proceeds with cooperative reasoning and pedagogy integrated.

3.1. A full-blown teacher–student mental representation

In previous sections, we mentioned that ToM is the prerequisite of CL but did not rigorously define what a mind is. Now, we elaborate on the mental representations in the CL framework. In the CL framework between two agents, the teacher and the student, there are six minds. The first pair of minds comprises the egocentric minds of the agents. The second pair comprises the agents' estimation of their partner's mind. These four minds, together with a common mind shared by the agents and the mind of an oracle (God)[†], make up six minds in total. Theoretically, the mutual reasoning in multiagent systems is recursive and can be infinite if no convergence exists [7,61–64]. For example, A knows fact e; B knows that A knows fact e; A knows that B knows that A knows fact e, and so on and so forth, *ad infinitum*. To avoid computational intractability, we only model one level of recursion, which is corroborated by the cognitive ability revealed in multiple human studies [65,66]. Another caveat is that the representation in this section does not include the dynamics of the minds—that is, how a mind updates during the interaction between agents. We defer that part of the discussion to Section 3.3. For now, let us just assume that they have a way to evolve given the inputs from the world and the partner. An agent's mind includes the following components:

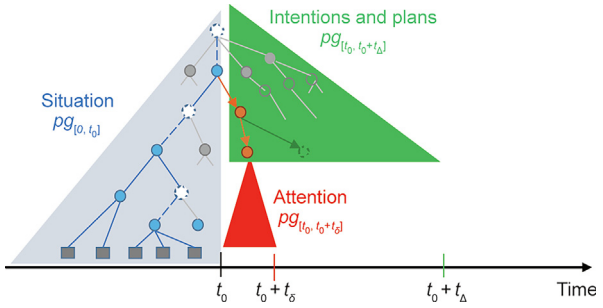

Fig. 6. State $w = pg$ unfolding over time. Here t_0 denotes current time, t_δ and t_Δ represent a short and a long elapse of time, respectively. $pg_{[0,t_0]}$: current situation; $pg_{[t_0,t_0+t_\delta]}$:current attention; $pg_{[t_0,t_0+t_\Delta]}$: intentions and plans.

(1) **A state w representing a situation.** Here, w stands for the world and is often structured as a mental state. For example, a state $w = pg$ is a parse graph and, in a more general case, w is a spatial, temporal, and causal (STC) parse graph, STC-pg that we have developed for vision [67,68], language [69,70], robotics [71], and commonsense reasoning [72–75]. The CL messages will be exchanged in a so-called situated communication—that is, the parse graph represents the composition of a scene, such as a living room, objects inside the scene, and actions happening together with causal changes of fluent (also known as the time-varying states of objects). The situation is unfolding over time; as Fig. 6 shows, a parse graph pg consists of three parts:

- $pg_{[0,t_0]}$ summarizes the current situation (shown in blue in Fig. 6);
- $pg_{[t_0,t_0+t_\Delta]}$ predicts intensions and plans (shown in green);
- $pg_{[t_0,t_0+t_\delta]}$ depicts the current attention (shown as a red triangle).

As CL is an iterative process, the parse graph will be communicated via messages over time at multiple semantic levels. CL agents have a common mind, whose state $w_c = pg_c$ (Fig. 7) also contains: ① the shared situation $pg_{c[0,t]}$; ② shared goal/intents $pg_{c[t,t+t_\Delta]}$; and ③ shared attention $pg_{c[t,t+t_\delta]}$. Such representation is key to human communication, as suggested by anthropology and cognitive studies [3], so that CL agents can "get to the point" rapidly at the right level of detail. Thus, in order to deal with complex situations, CL is a situated learning and communication process, which is more general than existing learning methods. Some prototype CL learning has been demonstrated in HRI and teaming [70,71,75].

(2) **A model $\theta \in \Theta$ in a hypothesis/model space Θ.** In a deterministic setting, such as Valiant's PAC learning, a model is a concept represented by a set, such as an object category, which is equivalent to a uniform distribution over this set. In a probabilistic setting, a model defines a probability distribution $p(s;\theta)$ on the state space, usually referring to the parameters of the distribution. It can be the hyperplane of a support vector machine (SVM), the weights of a deep neural network (DNN), or the rules of a stochastic grammar. When state $w = pg$ is a structured parse graph with varying configurations, we represent model $p(w;\theta)$ with an and–or graph (AOG) [67], whose language is a set of all valid configurations associated with probability. A parse graph pg is an instance and realization of the grammar or AOG. In its full-blown complexity, a model is represented by an STC–AOG integrating three hierarchies:

- **Spatial hierarchy:** scenes–objects–parts–primitive, for parsing scenes and objects;

[†] This mind is also considered to be the objective world or nature, whose dynamic is attributed to two factors: The first is a set of physical rules, either deterministic or stochastic, which is inherent to the world, and the second is agent actions.

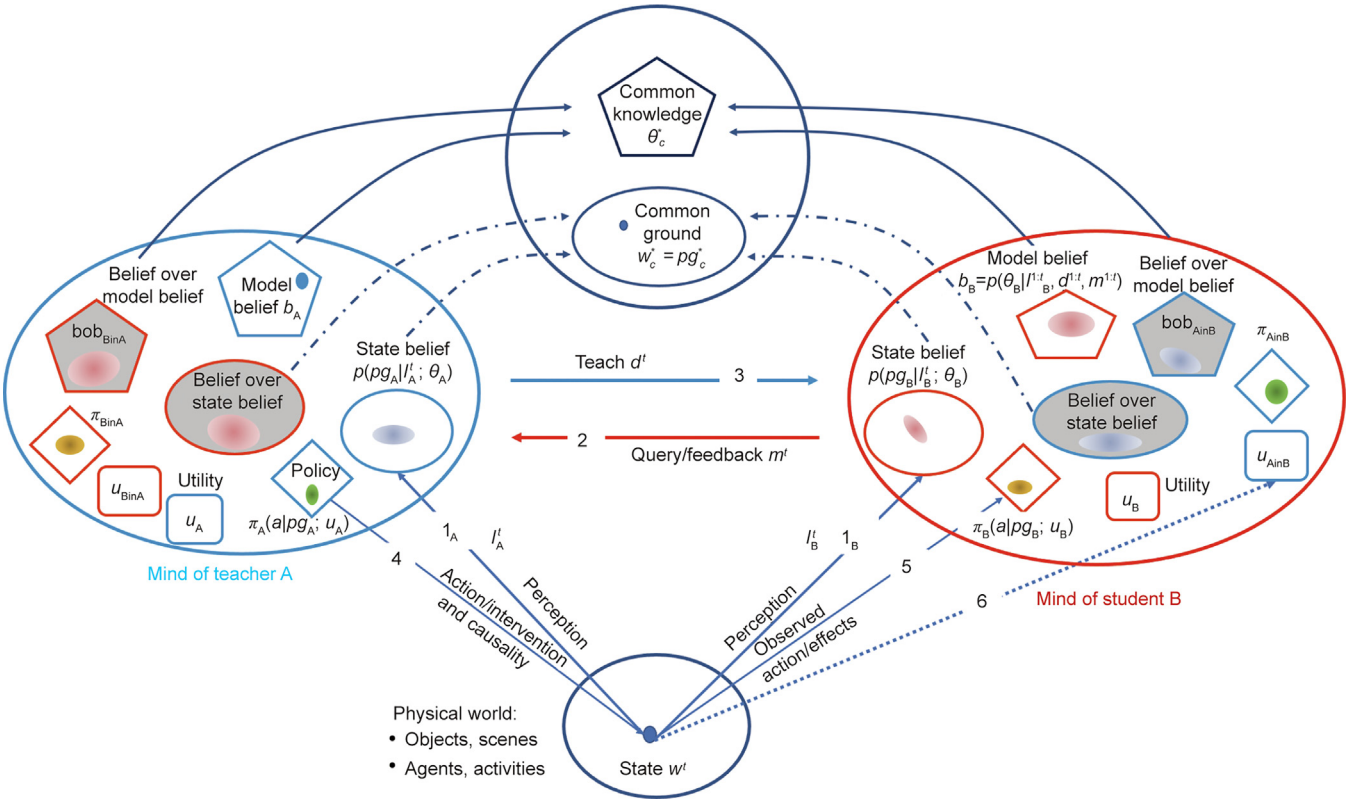

Fig. 7. A zoomed-in view of CL representations: unifying all existing learning protocols and beyond. Each mind contains four spaces: ① a pentagon representing the hypothesis/model space Θ; ② an ellipse representing the situation $w = pg$ in the state space W, where the beliefs are denoted by b and represented by clouds; ③ diamonds representing policy π, distributions over action a; and ④ squares representing utility function u. Arrows illustrate the dynamics: observation (I), intervention, and messages; shaded shapes represent belief over belief (BoB). Subscripts are used to indicate the teacher (A), the student (B), and their common knowledge (C). For recursive subscripts, the owner of the component comes after *in*. For example, π_{BinA} is in A, the teacher's mind. The parameters and variables in the figure are defined in Sections 3.1 and 3.2.

- **Temporal hierarchy and compositions:** events–actions–movements for parsing events;
- **Causal hierarchy:** actions and fluents for causal reasoning and task planning.

STC–AOG can be viewed as a unified representation for image parsing in computer vision, language parsing in natural language understanding (NLU), task planning in robotics, and cognitive reasoning in commonsense AI.

(3) **Belief and belief over belief (BoB).** We denote the teacher's observations and input by I_A, and the student's input by I_B. I_A and I_B can be an input image or video. We also denote the messages from the teacher at time t as d^t and those from the student as m^t. Also, let $a_{A/B}$ be the action of agent A/B . As neither the physical world nor the other agent's mind is fully observable to an agent, most of the time, agents need to form beliefs over the structure of interest. We denote the history for the student B up to time t as

$$h_B^t = \left[I_B^{1:t}, d^{1:t}, m^{1:t}, a_B^{1:t}\right] \tag{2}$$

where $d^{1:t}$ means $\left\{d^1, d^2, ..., d^t\right\}$ and $m^{1:t}$ means $\{m^1, m^2, ..., m^t\}$.

Then, B's belief of the state and the model are $b_{B,w}^t(w) = p\left(w|h_B^t\right)$ and $b_{B,\theta}^t(\theta) = p\left(\theta|h_B^t\right)$. We do the same for agent A, except that we usually assume that $b_{A,\theta}^t(\theta) = b_{A,\theta}^0(\theta)$; that is, we assume that the teacher has a correct and static model at the beginning. For the ToM structures, uncertainty also exists. Thus, BoB is defined as follows:

$$\text{bob}_{\text{AinB},\theta}^t(b_{A,\theta}) = p\left(b_{A,\theta}|h_B^t\right) \tag{3}$$

$$\text{bob}_{\text{AinB},w}^t(b_{A,w}) = p\left(b_{A,w}|h_B^t\right) \tag{4}$$

where $b_{A,w}^t$ and $b_{A,\theta}^t$ are A's belief of the state and the model at time t. $\text{bob}_{\text{AinB},w}^t$ and $\text{bob}_{\text{AinB},\theta}^t$, are B's estimation of teacher's beliefs at time t. Same definitions can be applied to the teacher's mental structures.

It might be noticed that the history increases exponentially as time goes by, and the model of BoB is theoretically intractable in general cases. In practice, for the history, multiple independence assumptions and inductive biases are indeed required to simplify the computation of the posteriors. For the nested belief, due to the deterministic belief update of Bayesian filters, if the state space is not too big, approximation techniques such as particle filters [62] can handle the calculation of BoB most of the time. However, a particle filter can reach its limitation when the state space is large or even uncountably infinite, a usual case for bob_θ. In those scenarios, one can continue the learning by taking the maximum a posteriori (MAP) estimation as the approximation of the distribution [53,76].

(4) **A decision policy** $\pi: W \mapsto \Delta(\mathscr{A})$[†] mapping the current state, w^t, to a distribution of actions $a \in \mathscr{A}$, where $\mathscr{A}$ is the action space. In more general cases, action can be composed into structured plans, such as in the form of T-*pg*. It should be noted that this action a

[†] We use $\Delta(X)$ here and in the rest of the article to represent the space of distributions over X.

refers to actions in executing real-world tasks, not how an agent sends messages in CL. The latter is a CL protocol, which we will elaborate later in Section 3.3. π_A and π_B denote the policy of two agents A and B. In practice, to represent such a mapping from states to actions, we usually assume that the policy has a certain form and only track its parameterization. Moreover, policies are usually determined by the agent's value, which is defined as the utility function.

(5) **A utility function** $u: W \mapsto \mathbb{R}$ represents the value of the agent—that is, what the agent cares about, the loss of mistakes, and the cost of actions. In combination with the model—that is, how the world transits given actions—utility functions can be used for task planning [71]. u_A and u_B denote the utility functions of two agents A and B. The agents A and B in CL must also estimate and learn the utility function of other agents, which we denote by u_{BinA} and u_{AinB}, respectively. As we will discuss later, u_{BinA} and u_{AinB} influence the CL protocol and equilibria of the learning process, resulting in different limits of learning.

3.2. A unifying framework of learning

To summarize the representation discussed in the previous section, Fig. 7 shows a zoomed-in view of the CL representations. There are six minds in total. Each agent has two minds colored differently—one for itself, and one for its partner. Each of these four minds has four components (w, θ, π, u), represented by different shapes. We model the uncertainty of w and θ by maintaining beliefs and BoB for them. The other two minds are the common mind and the physical world—that is, God's mind.

The arrows in Fig. 7 show the various dynamics and information flows, which include three types:

- Observations I_A and I_B from the physical state to the perceived state space W;
- Actions or interventions that cause changes in the physical state;
- Messages between the two agents to exchange information. Depending on the learning modes, these messages are for inference, learning, demonstration, confirmation, and so forth.

For clarity, we omit arrows for other dynamics; for example, some messages may be generated from a BoB space to probe what the other agent is thinking, such as, "I think your state estimate w is ..." or "what do you know about the state w?" Some arrows are second order; for example, the teacher learns the policy π_{BinA} from observing how the student conducts a task, that is, LfD [30,75], or learns the utility u_{BinA} by watching the student's decision or choice [71].

In CL, the communication between the teacher and the student converges at three levels (see the curved arrows in Fig. 7):

- When the inference process converges, they reach a common ground or situation w_c^*.
- When the learning process converges, they reach a common model knowledge θ_c^*.
- When their policy and utility converge, they reach a common social norm π_c^* and ethics u_c^*.

Depending on the learning protocols and characteristics (i.e., capacity for generating and interpreting messages) of the agents, the convergences may have different equilibria that decide the limits of learning. In CL, we assume that the agents are cooperative and not deceptive, and that their utility functions are aligned through learning. In most learning setups, convergence is considered in the second and the third level, and we will use them as the convergence criteria for most of this article. We will discuss multiple levels of convergence in Section 5.2.

The motivation of CL is to integrate the essence of human pedagogy into machine learning and thereby overcome the limitation of existing algorithms. In fact, with current representations, we can see that CL does unify the existing learning methods, with the following relations:

- Shannon's communication diagram in Fig. 3 is a message-passing channel between two agents. CL extends this communication setting by including the mental states, the BoB space, utility functions, and a common mind, all of which evolve over time. This allows more sophisticated messages and enables agents to "read between the lines."
- Valiant/Vapnik's theory is passive inductive statistical learning, supervised or unsupervised, from randomly sampled examples. This is shown by arrow 1_B in Fig. 7. In contrast, in CL, messages are deliberated based on a reflection of the mental states and utility functions.
- Active learning is represented by arrow 2: B can ask A to label certain examples selected by B. The example is selected to gain the most information in optimizing B's utility/loss function.
- Algorithmic teaching [77,78], shown as arrow 3, is a protocol complementary to active learning. Teacher A chooses the best examples to teach a student B for efficiency. A must consider what B knows and select critical examples to B, such as support vectors for classification.
- LfD [15,30,75] is a typical learning protocol in robotics and is an important component for commonsense acquisition. This learning method is shown by arrows 4 and 5 in Fig. 7. Agent A teaches a task by performing a sequence of actions on objects. The student observes the actions and their outcomes directly and learns the action policy from the teacher.
- Causal learning is represented by arrows 1_B and 5, where an agent applies actions to change the fluents of objects and scenes, and learns the causal effects of its action in terms of changed object fluents, including appearance changes (e.g., painting a wall, mopping a floor), geometry changes (e.g., blowing up a balloon), and topology changes (e.g., cutting a fruit).

The CL can also create new learning methods or protocols that are not well known. For example:

- Perceptual causality learning. This is in contrast to causal learning [79], where the experiment/intervention requests A to perform actions (arrow 4) and observes the effects of her actions (arrow 1_B and 5). In Ref. [73], a new protocol called perceptual causality learning is proposed. Here, the student can learn causality by watching (arrows 1_B and 5) the actions of the teacher (arrow 4). Under the assumption that she is not performing magic (i.e., no cheating), the student will infer and mirror the actions of the teacher in what is called "perceived causality." As shown in Ref. [73], this protocol is far more effective for learning causality and opens the door for learning causality from observations—a key aspect of human intelligence.
- Utility learning is shown by arrows 4 and 6. The student infers the utility function of A by observing her decisions and choices in actions. Economics theory says that rational agents make decisions and take action for utility maximization. By observing the actions taken by A, the student can infer A's utility, denoted by u_{AinB} in CL. For example, in the example about folding a T-shirt demonstrated in Ref. [71], by watching the teacher folding T-shirts, the student can not only learn the causality and policy π_A but also learn a utility function u_{AinB} for aesthetics—that is, what states of the T-shirt have a relatively high value to the teacher. B may choose to adopt a similar utility function $u_B \leftarrow u_{\text{AinB}}$. In CL, agents will update and align to a common utility. We will later show another value alignment example in Section 4.3.

- Learning by analogy (LbA) is a powerful learning mode that is used by humans [4] but is missing in current popular machine learning methods. It requests shared knowledge (w_c, θ_c) between two agents and the capabilities of abstraction and projection to transfer knowledge across domains using an abstract graphical representation. Abstraction and projection are key intelligent capabilities in classic Raven's intelligence quotient (IQ) tests but are missing in current statistical learning. The shared mind will facilitate LbA. As the common mind grows, the two agents will be more and more synced and the student will become as capable as the teacher.

To summarize, all learning paradigms above having their most suitable using scenarios, and meanwhile, they illustrate the usage of various special cases of the CL framework. Nevertheless, the deployment of general human-like AI usually involves much more complicated and comprehensive settings, in which multiple components of the full CL framework become indispensable. In Section 4.3, we show an example scenario bearing adequate complexity to summon the full structure. Now that we have wrapped up the introduction of the CL representation, we can move on to the specific formalism that drives the learning and models the mind update dynamics.

3.3. Formalism of CL

In Section 3.1, we introduced the representation of the CL framework, with which the integration of human pedagogy and ToM with machine learning becomes possible. In this section, we show the dynamics of the teacher's and the student's minds. This completes the CL formalism and enables us to scrutinize prior machine learning algorithms by instantiating this generic formalism.

3.3.1. Overall setting

First, let us recall that a standard machine learning algorithm is a mapping from training data to a model space [18,40], where a model can be, for example, a specific hyperplane in SVM, the location of the kth centroids in K-means, or the parameters of a neural network. In CL, we generalize learning into a pedagogy process involving two agents—a teacher and a student. Each agent has its model; or, in cases where uncertainty must be considered, a belief of the model. In Section 3.1, we differentiate an agent's model with its policy and utility to better specify various types of learning in a situated communication scenario. Nevertheless, for modeling the learning dynamics, we do not need to differentiate them explicitly. That is, the model that we are referring to here is a representation with broader meanings than the model in Section 3.1. The only purpose of the latter is to interpret the physical world. In Section 3.4 and Section S3, we show how concept, policy, value, and utility can all be learned as the model with the same CL formalism.

Let us rigorously denote the teacher's model space as Θ and $b_\theta \in \Delta(\theta)$ as the teacher's belief of the model[†]. We assume that b_θ stays constant across the pedagogy process, because the teacher will not receive any new information helping her to refine the model. Moreover, in most cases, we assume that the teacher knows the true model θ^*. That is, $b_\theta(\theta)$ becomes 1 $(\theta = \theta^*)$. Similarly, we have the student's model space Ω and the student's belief of the model as $b_\omega \in \Delta(\Omega)$. Notice that we do not assume that the teacher and the student have the same model space; that is, Ω may or may not be the same as Θ. Having two separate model spaces of the agents allows CL to deal with the teacher and student pairs having distinctive world representations. For example, suppose that the model is a policy network mapping from the robot's camera inputs to motion commands. Then, robots with different camera configurations are still able to teach and learn from each other. Since the student will update his belief of the model, we use b_ω^t to represent his belief at time t. The same superscript will be applied to other time-variant variables.

The goal of the CL is for the student to have an accurate enough belief of the model so that he can achieve a certain level of performance for a given task compared with the teacher. We can define this metric as minimizing a loss function $L\left(b_\theta, b_\omega^T\right)$, where T is the moment where the learning process terminates, and L measures the difference between the teacher's performance and the student's performance of a task. The smaller the performance gap is, the smaller L will be. At this moment, to define the framework of CL, we do not specify L in too much detail. In Section S3, we describe how different machine learning paradigms are special cases of CL and further concretize L.

3.3.2. Teacher setting

CL is a pedagogy process in which the student's belief of the model is refined, given messages from the teacher. At time t, the teacher chooses message d^t from her message space, $\mathscr{D}$. The message selection criteria at each time depend on the student's learning state at that moment, which is denoted as $s^t \in S$. The learning status is a variable maintained by the teacher to keep track of the student's progress at a particular time. For example, s^t can be the validation error at the time t. When the exact learning state is not directly available to the teacher, she needs to have a belief, $b_s \in \Delta(S)$, over the learning states. Just like L, we do not further restrain the representation of $\mathscr{D}$ and S until we examine specific learning paradigms.

In general, when the teacher has an accurate estimation of the student's learning state, the pedagogy can be more effective. Hence, the teacher has a transition model for the learning states: Namely, how the student will make progress after receiving a message. Assuming Markovian learning states, the transition model can be mathematically defined as follows:

$$\psi : S \times \mathscr{D} \mapsto \Delta(S) \tag{5}$$

That is, ψ takes in the student's current learning states and the teacher's message and returns the distribution of the student's new learning states. In situations such as active learning [29], the teacher also receives messages (query data) from the student. We can denote the student's message at time t as $m^t \in \mathscr{M}$. Without loss of generality, for the rest of the paper, we assume that m^t comes after d^t for every time t. Messages from the student can also help the teacher to estimate the student's current learning state. To accommodate these messages, the teacher has a message model for the student, mapping learning states to a distribution over the student's messages:

$$\phi : S \mapsto \Delta(\mathscr{M}) \tag{6}$$

With ψ and ϕ, the teacher can update her b_s condition on her outgoing and incoming messages using the Bayesian filter:

$$b_s^t(s) = p\left(s|d^{1:t}, m^{1:t}\right) \propto \phi\left(m^t|s\right) \int_{s' \in S} \psi\left(s|s', d^t\right) b_s^{t-1}(s')\,\mathrm{d}s' \tag{7}$$

In general, this belief cannot be calculated exactly, especially when $|S|$ is large or infinite. However, in practice, ψ and ϕ are usually modeled as indicator functions, effectively simplifying the computation. With the student's current learning state and the teacher's belief of the model b_θ, we can have a teaching policy for the teacher:

[†] In the rest of the article, the subscripts of beliefs are used as labels to differentiate different beliefs and are not to be confused with parameters, which present inside parentheses, unless explicitly defined.

$$p\left(d^t|b_\theta, b_s^{t-1}\right) = \text{softmax}_\beta\left(Q_\psi\left(b_\theta, b_s^{t-1}, d^t\right)\right) \tag{8}$$

where $\text{softmax}_\beta(x) = \frac{\exp(\beta x)}{\sum_{x' \in X}\exp(\beta x')}$ is the Boltzmann rationality [80–82] and $Q_\psi\left(b_\theta, b_s^{t-1}, d^t\right)$ is a value function that takes in the teacher's belief of the student's current learning state, the teacher's model, and a teacher's message. Usually, it can be further expanded into an integral form with a simpler defined $Q_\psi(\theta, s, d)$ weighted by $b_\theta(\theta)$ and $b_s^{t-1}(s)$. As this function can depend on the teacher's transition model of the student's learning state, it is parameterized by ψ. Q can either be learned from data or defined by experts [12,83]; in Section S3, we provide examples of both.

3.3.3. Student setting

In this section, we define the student in CL. Unlike standard machine learning algorithm mapping from data to a model, the student in CL is aware of the cooperative teacher in the pedagogy. That is, the student knows that he receives selected messages instead of random examples from the teacher. To model the teacher-aware student, we start with a teacher-unaware student, whose belief update rule given a new message d^t is simply:

$$b_\omega^t(\omega) = p\left(\omega|d^{1:t}\right) \propto b_\omega^{t-1}(\omega)\pi\left(d^t|\omega\right) \tag{9}$$

where $\pi{:}\Omega \mapsto \Delta(\mathscr{D})$ is the teaching likelihood function in the student's mind. In general cases, such an estimation of the teaching function does not match the exact pedagogy policy of the teacher defined in Eq. (8). However, as we will see in the coming sections, most of the time, when the student has a reasonable approximation, learning is effective.

Next, for a teacher-aware student, messages are conditioned on more than just the teacher's model. Using all the information available to him, a teacher-aware student has

$$b_\omega^t(\omega) \propto b_\omega^{t-1}(\omega)\pi\left(d^t|\omega, d^{1:t-1}, m^{1:t-1}\right) \tag{10}$$

The disadvantage of using the entire history in π is the exponential growth of the history with respect to time. Recall that the teacher relies on two things when teaching: One is her model, θ^*, and the other is her estimation of the learning state of the student, s. Thus, we define $\hat{s} \in \widehat{S}$ as the student's estimation of his learning state in the teacher's mind. Then, the full history, $d^{1:t-1}$, $m^{1:t-1}$ can be condensed into $b_{\hat{s}}^{t-1}$, and Eq. (10) becomes

$$b_\omega^t(\omega) \propto b_\omega^{t-1}(\omega)\pi\left(d^t|\omega, b_{\hat{s}}^{t-1}\right) \tag{11}$$

where $\pi : \Omega \times \Delta\left(\widehat{S}\right) \mapsto \Delta(\mathscr{D})$ is the teaching likelihood function accommodating the student's learning state in the teacher's mind. That being said, in every time step, the student maintains two beliefs: one for the model and one for the estimation of the teacher's impression of him. Theoretically, the nested mutual reasoning between the teacher and the student can be infinitely recursive [62]. To avoid this intractability and complexity, in CL, we stop at a teacher-aware student and do not go deeper into the recursion. To update $b_{\hat{s}}$, the student needs two more functions to model the transition of $\hat{s}$ after two types of messages. We define

$$\zeta : \widehat{S} \times \mathscr{D} \mapsto \Delta\left(\widehat{S}\right) \tag{12}$$

$$\xi : \widehat{S} \times \mathscr{M} \mapsto \Delta\left(\widehat{S}\right) \tag{13}$$

As the student's transition functions for $\hat{s}$—namely, how the teacher's impression of him will change after she sends (ζ) and receives (ξ), a message. The student's counterpart of ϕ should be a function mapping from $\hat{s} \times \Omega$ to $\Delta(\mathscr{D})$ as his estimation of how the teacher will teach given a model when she has learning state $\hat{s}$. This can be accomplished by π, because any $\hat{s}'$ can be written as a Dirac-delta distribution, $\delta_{\hat{s}'}(\hat{s}) = 1$ $(\hat{s} = \hat{s}')$. Notice that ζ, ξ, and π are all approximations of the teacher's mental change in the student's mind, so Ω and $\widehat{S}$ are used instead of Θ and S.

Now, we can write down the belief update function for $b_{\hat{s}}^t$. Unlike Eq. (7), we want to have an intermediate variable, after the student receives d^t and before m^t is sent out, whose purpose is to determine the best m^t to send out. Let us denote

$$\begin{aligned}\widetilde{b}_{\hat{s}}^t = p\left(\hat{s}^t|d^{1:t}, m^{1:t-1}\right) \approx \frac{1}{Z}\int_{\omega\in\Omega,\,\hat{s}\in\widehat{S}} \zeta\left(\hat{s}^t|\hat{s}^{t-1}, d^t\right) \\ \times \pi\left(d^t|\delta_{\hat{s}^{t-1}}, \omega\right) b_{\hat{s}}^{t-1}(\hat{s}) b_\omega^t(\omega)\mathrm{d}\omega\mathrm{d}\hat{s}\end{aligned} \tag{14}$$

as the student's belief of the teacher's estimation of his learning state after he receives d^t, where Z is a normalizing factor. A detailed derivation can be found in Section S1. Using $\widetilde{b}_{\hat{s}}^t$, the student comes up with the message to send to the teacher, with

$$p\left(m^t|\widetilde{b}_{\hat{s}}^t, b_\omega^t\right) = \text{softmax}_k\left(V_\xi\left(\widetilde{b}_{\hat{s}}^t, b_\omega^t, m^t\right)\right) \tag{15}$$

where $V_\xi\left(\widetilde{b}_{\hat{s}}^t, b_\omega^t, m^t\right)$ is a value function for the student. Here, V takes the belief as its argument, because the student usually needs to take the attributes of the distribution, such as the entropy, into consideration for message selection. Just like the value function, Q, to the teacher, V can either be learned from data or defined.

After the student gives his message m^t to the teacher, he finalizes his belief of $\hat{s}$ for time t using

$$b_{\hat{s}}^t(\hat{s}) = p\left(\hat{s}|d^{1:t}, m^{1:t}\right) \propto \int_{\hat{s}'\in\widehat{S}} \xi\left(\hat{s}|\hat{s}', m^t\right)\widetilde{b}_{\hat{s}}^t(\hat{s}')\mathrm{d}\hat{s}' \tag{16}$$

Eq. (16) concludes all the belief updates in CL. As it involves many components and nested inferences between agents, we summarize the CL framework with Fig. 8 and specify it in Algorithm 1. In the next section, we demonstrate how various learning paradigms can be expressed as special cases of CL.

Algorithm 1: Communicative learning.

Input-teacher: Θ, b_θ, b_s^0, Q, ψ, ϕ, $\mathscr{D}$
Input-student: Ω, b_ω^0, $b_{\hat{s}}^0$, V, ζ, ξ, π, $\mathscr{M}$
Input-world: T, L
Output: b_ω^T

For $t = 1{:}T$ **do**
  Teacher selects d^t using Eq. (8)
  Student updates $\widetilde{b}_{\hat{s}}^t$ and b_ω^t using Eqs. (14) and (11)
  Student selects m^t using Eq. (15)
  Teacher updates b_s^t using Eq. (7)
  Student updates $b_{\hat{s}}^t$ using Eq. (16)
End
Return b_ω^T, $L\left(b_\theta, b_\omega^T\right)$

3.4. CL instantiation of prior learning paradigms

The formalism proposed in Section 3.3 gives us a unifying lens to summarize existing learning paradigms. That is, we can instantiate the CL formalism with different learning paradigms by constructing their corresponding value and partner estimating

functions. Since not all of these paradigms include the full set of recursive mutual reasoning in CL, we can classify them according to their recursive level. If a learning paradigm only has a student, and data comes from a random process instead of a cooperative teacher, we call it a level-0 paradigm. Some level-0 paradigms can involve multiple agents, such as the co-training/teaching [84–86] framework for label-noisy learning. In spite of having two learning networks, it still follows a variation of passive learning, because there is no mutual reasoning between the learning agents, which only differ by parameter initialization and seek to maintain such divergence during training to avoid confirmation bias. Paradigms like active learning [29], where the teacher only passively replies to the student's queries about random data and has no agency, are also classified as level-0. Suppose that there is a cooperative teacher who can choose her messages to the student, but the student is not aware of the existence of this teacher. Such learning paradigms are termed as level 1. When the teacher is cooperative in message selection and the student is aware of her helpfulness, we classify it as a level-2 paradigm.

We summarize the paradigms and their level in Table 1. It can be seen that all these learning paradigms are essentially special cases of CL with one or a few components omitted from their modeling. Within each learning paradigm, there can be various learning algorithms, and every algorithm maps to a grounding of the CL components. In Section S3, we select some algorithms for each paradigm and show their CL groundings. As the purpose of this article is to propose a unified framework of learning rather than to exhaustively list all learning algorithms, we only choose a few exemplar algorithms for each paradigm and specify their CL groundings. For the other algorithms, their CL groundings can be easily migrated from the exemplar algorithms belonging to the same paradigm. The complete summary and comparison of these groundings can be found in Tables S2 and S3 in Appendix A.

One follow-up question about the CL formalism is how to acquire the belief-update and value functions. There are two usual approaches: The first is to use predefined heuristic functions, which are specifically designed or engineered for particular tasks. Most learning algorithms in Section S3 and the HRI example in Section 4.3 fall into this category. The second approach is to learn these functions as an emerged norm of communication. Learning to teach [42] (detailed analysis in Algorithm S10 in Appendix A) and the referential game example in Section 4.1 belong to this category. That is, after learning and teaching multiple models (every model learning follows Algorithm 1), the teacher and the student start to know each other better and form a tacit norm of communication. The convergence of the model and the convergence of communication norms form different hierarchies in learning. In Section 5.2.1, we return to this problem in more detail within the context of the halting problem of learning. At present, let us scrutinize CL with some exemplar applications. Using these examples, we will discover how the belief-update and value functions are defined or learned.

Table 1
Learning paradigms in prior work and their recursive level.

Paradigm	Teacher	Student	Level
Passive learning	None	Unaware	0
Active learning	Oracle	Unaware	0
LfD	Expert	Unaware	0
Algorithmic teaching	Cooperative	Unaware	1
CL	Cooperative	Aware	2

4. Examples of CL

Thus far, we have completed the definition of the CL formalism. In this section, we reify CL with a few examples. First, we demonstrate the effectiveness of a pragmatic protocol that emerges from applying the CL formalism to referential games [87]. Then, we quantitatively justify the advantage of CL with the empirical results of a teacher-aware learning algorithm on multiple benchmarks. Finally, we present a usage of CL in a realistic HRI task.

4.1. Case study 1: The emergence of a pragmatic protocol in referential games

In this section, we use a proof-of-concept referential game as an example to demonstrate the usage of the CL formalism. Despite its simplicity, the referential game encompasses the necessary components of standard communication between collaborative agents. It has been shown that effective communication protocols can emerge between agents playing various forms of referential games [88,89]; however, in this example, we illustrate how the CL formalism can facilitate the emergence of a more efficient protocol, which we term the pragmatic protocol. We also show how the belief-update and value functions defined in Section 3.3 can be learned via cooperative interactions supervised with the outcomes of referential games.

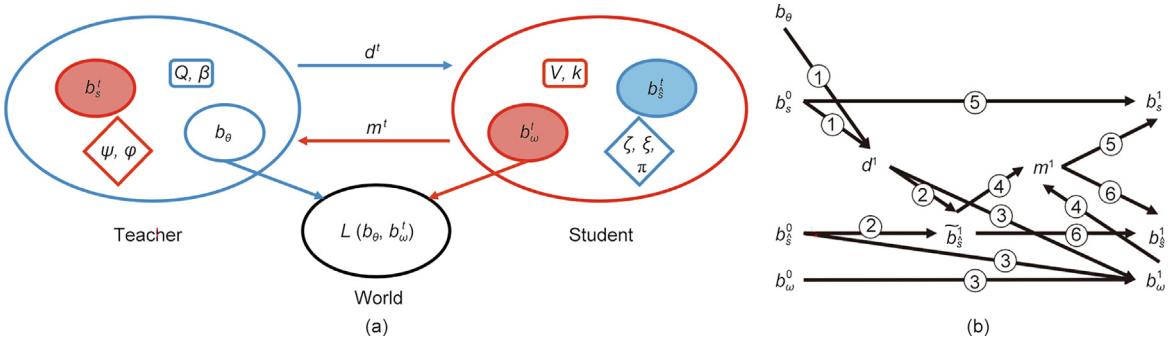

Fig. 8. CL formalism. (a) Mental state representation of the teacher and the student. The blue bubbles are for the teacher, and the red ones are for the student. Beliefs are drawn as ellipses, with the time-variant beliefs shaded. Diamonds hold the functions that agents use to maintain their beliefs about their partner's behavior and knowledge. We call these the belief-update functions. The value functions for the teacher and the student are in the blue and red rectangles. (b) The temporal structure of CL. This expands the transition process at the first time step. The numbers group the arrows, and the same number indicates that operations happen in one function. The "1" arrows correspond to Eq. (8), the "2" arrows correspond to Eq. (14), the "3" arrows correspond to Eq. (11), and the "4" arrows correspond to Eq. (15). Eqs. (7) and (16) are represented by the "5" and "6" arrows, respectively. The temporal order of these belief updates is the same as the numerical order of the arrows, with the "2"s and "3"s happening concurrently and the "5"s and "6"s happening concurrently.

4.1.1. Background: Referential game

The example we mentioned in Section 2.3 is actually a referential game. There are two agents participating the referential game, a teacher and a student. Every game includes a target object and a few distractors. Only the teacher knows the target, and she aims to notify the student with a message so that the student can identify the target out of the distractors after receiving this message. The infrastructure of CL enables ToM: In order to establish proper communication, the teacher and the student must take their partner's perspective into account [83]. Fig. 9 presents an example. There are three candidates, a blue sphere, a red sphere, and a blue cone, with the blue sphere being the target. If we only allow the messages to be colors and shapes, then, for an unaware student, the blue sphere cannot be uniquely identified, because both "blue" and "sphere" are consistent with multiple candidates. However, a teacher-aware student, after receiving "blue" from the teacher and conducting pragmatic reasoning, should be able to rule out the blue cone and identify the blue sphere. He knows that, if the target is the blue cone, a helpful teacher would have used "cone" to refer to it unambiguously. This example illustrates that, during pragmatic [10] communication, a message conveys more information than its literal meaning. The selection and disuse of certain messages can also suggest the intention of the speaker.

4.1.2. Emergence of a pragmatic protocol

Pragmatics studies the contribution of context to language meanings. In human communication, interpretations of language never take place out of context, and sentences can usually convey information that goes beyond their literal meanings. However, this mechanism is missing in most multiagent systems, restricting the communication efficiency and the capability of human–agent interaction. The example in Fig. 9 concretizes a typical pragmatic principle called the scalar implicature. More details will be discussed in Section 5.1.2. It is only when agents have ToM and are able to capture their partners' cooperative intention that they can bear pragmatic reasoning. Luckily, the infrastructure of CL supports ToM and allows the emergence of pragmatics protocols.

4.1.2.1. Overview. Before speaking, the teacher traverses all messages and predicts the student's new belief after receiving each message using ψ. She then sends the message leading to the most optimal student's new belief. Hearing the message, the student updates his belief and takes action (makes a decision or waits for another message). The recursive mutual modeling in ToM is integrated within the belief update process. The beliefs in our model are semantically meaningful hidden variables in the teacher's Q-function as defined in Eq. (3), and the student directly samples an action according to his belief. The evolving of the belief-update function ψ and π reflects the protocol dynamics between the agents.

4.1.2.2. Teacher. The teacher selects messages according to her Q-values and belief-update function ψ, which takes in the candidate set, current belief estimation, and a message. The return value of this function is a new belief estimation. This function can be parameterized as a neural network with weighted candidates encoding and messages as the inputs and softmax as the output layer. The return value of the belief update function is directly fed into the Q-function. That is, the output of the belief update function is used in Q to predict the student's belief in the next step during testing. The teacher chooses messages according to Eq. (3).

To form a protocol, the teacher needs to learn two functions: ψ and Q. In the training phase, every time the student receives a message, he returns his new belief b^t_ω to the teacher. During testing, she needs to use the output of ψ to approximate the student's new belief. We train ψ by minimizing the cross-entropy between b^t_ω and the teacher's prediction. Teacher's Q is learned with Q-learning [90], given the reward as the results of the referential game.

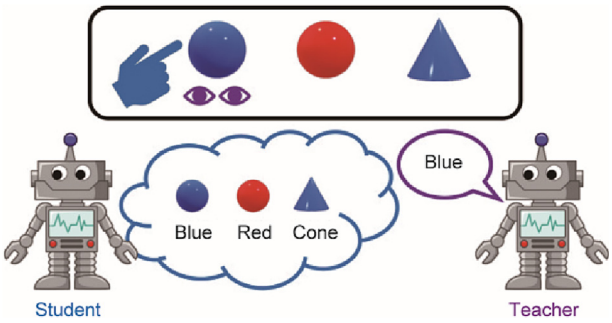

Fig. 9. An example referential game. From right to left, the candidates are a blue cone, a red sphere, and a blue sphere. If a student receives "blue" from the teacher, he should be able to identify the blue sphere instead of the blue cone.

4.1.2.3. Student. In the referential game setting, the student does not need to give much feedback. Thus, we only introduce his belief-update function π and policy of making decisions, which triggers the game reward for the norm of communication to emerge within the group. We directly learn π and the policy of the student through the REINFORCE algorithm [91]. In the referential game, the student's policy is quite simple. If his belief is certain enough, he will choose the target based on his belief; otherwise, he will wait for further messages. π and the policy can be parameterized as an end-to-end trainable neural network, with the candidates' encoding, original belief, and the received message as inputs and returning an action distribution.

4.1.2.4. Adaptive training. Both the teacher and student are adaptively trained to maximize their expected referring success. We initialize the student as unaware, that is, with π following a Bayesian belief update process. Then, we fix the student and train the teacher to update her Q and ψ. After a fixed amount of time or until converging performance, we fix the teacher and train the student. The alternate fixing and training continue until the group performance no longer improves. The final pragmatic protocol will be fulfilled by then.

4.1.3. Protocol analysis

The emerged protocol demonstrates pragmatic reasoning, which results in a significantly better referring performance than other emerged protocols [87]. In Fig. 10, we show some example referential games together with the teacher's and the student's beliefs during training. It is clear that the student can differentiate targets from distractors, even when messages from the teacher have consistent literal meanings with all of them. That is, both the literal and the pragmatic meanings of messages are correctly grasped. More specifically, if the student is shown the four three-dimensional (3D) objects in Fig. 10, he will know that the teacher will send "blue" for the blue ellipsoid, "large" for the large green cylinder, and "magenta" for the magenta cylinder. Hence, "upper right" and "ellipsoid", despite being valid descriptors for multiple candidates, must indicate the target green ellipsoid on the upper right. Mathematically, the emerged pragmatic protocol approximates the RTD of the candidate concept space, which measures the number of examples needed for concept learning between a pair of cooperative and rational agents [87,92]. The formal definition of RTD is provided in Section S4 in Appendix A. Intuitively, in a concept class, there is a subset of concepts that are the

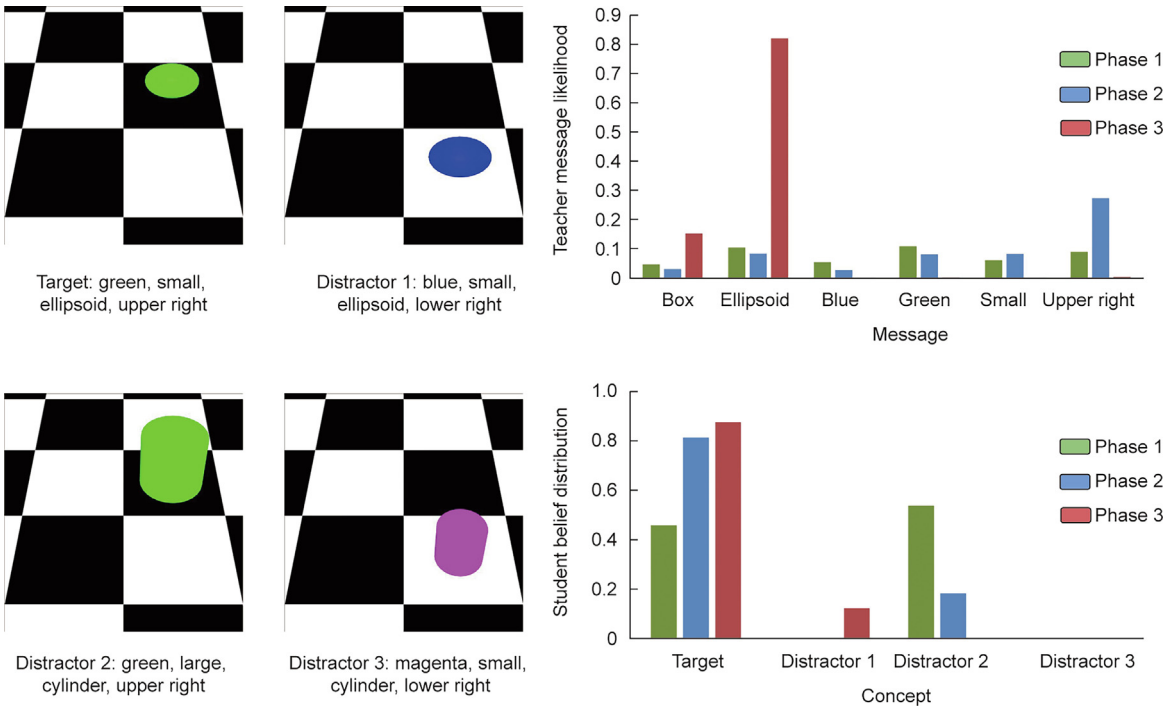

Fig. 10. A referential game example with four candidates. The fixing and training roles of the teacher and the student are switched at the end of every phase. We show the message distribution for the target and the student's new beliefs after receiving the most probable message. The teacher concentrates all her message distribution weights on the unique identifiers after the first phase of training. The student's belief illustrates that the teacher's most probable message—although consistent with multiple candidates—can successfully indicate the target with more confidence as the training proceeds. In general, both agents' behavior becomes more certain, and the certainty coordinates.

simplest to learn—that is, that have the minimum-sized teaching set among all the concepts. It is possible to first learn those concepts and remove them from the concept class. Now, for the remaining set of concepts, it is possible to recursively learn the simplest concepts, and so on. The teaching complexity of this learning schema lower bounds classic teaching dimension (TD) [93]. In every phase of our iterative training, the agent learns to identify the optimal teaching set for the "simplest" remaining candidates. In our case, candidates that are identifiable with a unique message are the simplest.

To summarize, the referential game example illustrates how the CL formalism can be used to develop new learning protocols from scratch with only signals about communication outcomes. In the coming sections, we provide exemplar usages of CL in more realistic and more complicated settings than referential games to mark the generality and scalability of the formalism.

4.2. Case study 2: Iterative teacher-aware learning

In a referential game, the teacher aims to indicate the target among distractors to the student. This process can be formulated as a concept learning problem with a discrete concept space [18,94]. The tractability or even finitude of the concept space makes referential games and similar concept learning problems appropriate starting points to study the advantage of CL [45,55]. However, to cover the full spectrum of machine learning paradigms, CL must accommodate problems with intractable hypothesis spaces, such as learning continuous parameters [38–41].

In this section, we examine the theoretical convergence guarantee of CL using machine parameter learning as an example. We compare two types of learning paradigms. The first involves a cooperative machine teacher and a naive learner; the second involves the same teacher and a teacher-aware learner [95]. The learner estimates the teacher's data selection process with distribution and corrects his likelihood function with this estimation to accommodate the teacher's intention. Maximizing the new likelihood enables the learner to utilize both explicit information from the selected data and implicit information suggested by the pedagogical context. It is clear that the teacher and student group following the CL formalism can achieve better performance both empirically and theoretically.

4.2.1. Background: Machine parameter teaching

Due to its continuous state space and long horizon planning, optimal parameter teaching has been a challenging problem. Thus far, the most prevailing and promising framework is machine teaching [39,40]. Thus, we adopt an iterative variation of machine teaching [41]. Following the notations in Section 3.3, the teacher holds $b_\theta(\theta) = 1\ (\theta = \theta^*)$, where θ^* is the fixed ground-truth parameter only known by the teacher. θ^* can be acquired by minimizing a loss function defined on data following a distribution, such as the mean squared error for regression, cross-entropy loss for classification, and negative log-likelihood for inverse reinforcement learning (IRL) [96,97]. In this section, we assume the choice of the loss function is the teacher and student's common knowledge.

4.2.1.1. Overview. The teacher and the student can represent the same data example in different but deterministically related ways. Then, it is likely that the optimal parameters for the teacher and the student are in different spaces too. This mimics practical scenarios: The teacher and the learner are a human and a robot, or two robots manufactured differently. Hence, following the notation in Section 3.3, we use $\theta^* \in \Theta$ and $\omega \in \Omega$ to indicate the parameter for the teacher and the student, respectively. As the representation of examples can be complex, such as features extracted by DNNs [25,27], without harming the expressiveness, we assume that the final decision is made by applying a linear model to the data representation.

To avoid infinite mutual reasoning between the teacher and the learner, we restrict the mutual knowledge between them. In this section, we consider a teacher who assumes that the naive learner is using stochastic gradient descent (SGD) to update his model. Meanwhile, the learner knows that the data comes from a helpful teacher instead of being random. That is, the naive learner, the teacher, and the teacher-aware learner have level-0, level-1, and level-2 recursive reasoning, respectively, as defined in Section 3.4. These levels of recursion imitate human cognitive capability [65,66,98] and were also adopted in Ref. [53].

4.2.1.2. Teacher. Following the standard setup of machine teaching, the teacher and the student can only communicate via examples. This limitation does not impact the generality of the framework, as the examples can have generic formats, such as demonstration used in the IRL [96,97,99,100]. As defined in Section 3.3, data d^t are provided iteratively. The teacher's goal is to provide helpful examples to the student so that his parameter ω converges to its optimum ω^* as quickly as possible. Since the teacher cannot access ω^t or ω^*, the student gives some feedback m^t to her in each iteration to keep her updated with the pedagogy progress.

4.2.1.3. Naive learner. The teacher-unaware learner uses a simple learning algorithm, such as SGD [41,42,101,102] for iterative gradient-based optimization.

4.2.2. Iterative teacher-aware learner

We first delineate the teacher whom the student should be aware of. Similar to her counterparts in referential games, the teacher in parameter learning selects messages according to her Q-values and belief-update functions. In Section 4.1.2, we described how these functions can be learned by letting the teacher and the student play referential games. Due to the larger parameter space and long horizon planning, the teacher uses a greedy heuristic to define the Q-function—namely, the example that decreases the distance between the student's parameter and the ground-truth parameter the most, a criterion commonly used in machine teaching problems [39–41,53]. The improvement brought by the example is defined as its teaching volume [41] and used as its Q-value, whose mathematical definition can be found in Eq. (S40) in Section S3.2 in Appendix A. Since the teacher does not have access to the student's current parameter in most practical settings and, even if she knows the parameter, she cannot utilize the value in a different parameter space directly, we let the student return the inner products of his parameter and the data to the teacher as feedback [41,95]. With the linear model assumption, the teacher can capture the student's learning states using simple belief-update functions. See the row of iterative machine teaching (IMT) [41] in Appendix A Table S2 for details.

IMT proposes a cooperative machine teacher that significantly helps the naive learner but still falls short of complete pedagogy by neglecting the student's teacher-awareness. To fill in this gap, we took advantage of CL and proposed iterative teacher aware learning (ITAL) [95]. The teacher-aware student adjusts his π function by taking the teacher's example selection into consideration. Intuitively, given all data in a mini-batch, the teacher chooses one specific example but not others. With what parameter can the probability of this selection be maximized? To this end, the updated teaching likelihood function has two components: The first models the consistency between the parameter and the data (the literal meaning of the example); the second models the teacher's selection with an estimated teaching volume and counterfactual reasoning (the pragmatic [10] meaning of the example). With the help of the new π, ITAL achieves both theoretical and empirical improvements on various regression, classification, and IRL benchmarks [95]. Fig. 11 [95] compares the learning curves of the teacher-aware learner against those of the naive learner.

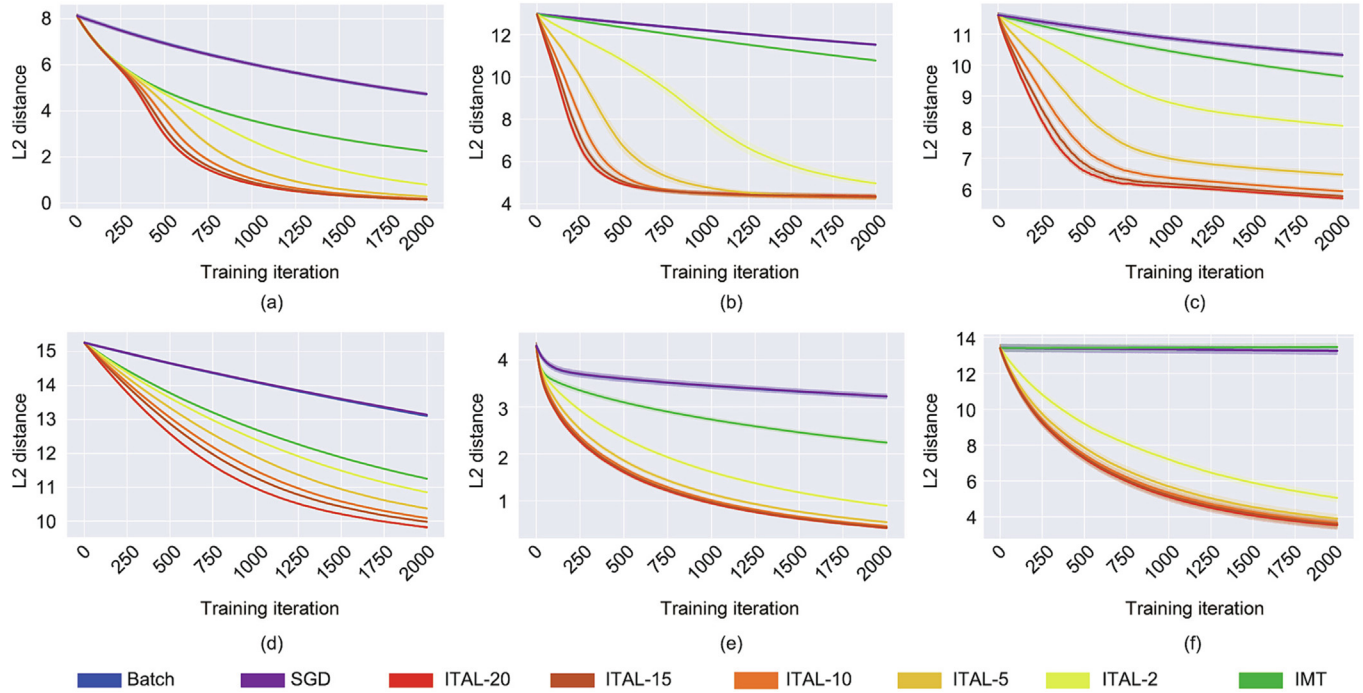

Fig. 11. ITAL performance on different tasks. (a) Linear regression; (b) Gaussian data; (c) CIFAR-10; (d) tiny ImageNet; (e) equation; (f) online IRL. With the same cooperative teacher, ITAL always achieves a substantial performance gain over IMT, demonstrating the benefit of teacher-awareness brought by CL. Within 2000 steps, ITAL already shows convergence, while a naive learner only learns to a limited extent in most tasks. The L2 distance is calculated using the ground-truth parameter $\omega^* \in \Omega$ and the student's current parameter. Batch and SGD teach a naive learner with random sampled data. ITAL-X represents the size of the mini-batch used by the teacher-aware student to estimate the teacher-volume. More details can be found in Ref. [95].

ITAL not only justifies the practicability of CL but also provides insights into generic human–machine teaming, because fast parameter learning enables machines to adapt to the user's needs quickly, even in real time. In the next section, we present the usage of CL in a human–robot collaboration setting.

4.3. Case study 3: Bidirectional human–robot value alignment

The overarching purpose of CL is to fulfill human-like learning ability and thereby achieve generic human–machine teaming. The machine would adopt the human user's input and change its behavior in real time so that the system and the human user can cooperatively accomplish a common task. To do so, the machine needs to actively infer the human user's belief, desire, and goal [103,104]. This inference process can be naturally modeled as a learning problem and fits into the CL formalism. In this section, we introduce a human–machine collaborative game in which a team of machine scouts must constantly align their value to the human commander's value to finish certain tasks [105]. We then demonstrate how the CL formalism can be applied to this scenario. As we will show, the success of the game completely relies on effective communications between the commander and the scouts, enabled by the representations in Section 3.1 and modeling in Section 3.3.

4.3.1. Communication paradigm for generic human–robot teaming

To achieve generic human-robot teaming, robots must be able to adapt to their users' values and change their behaviors in real time so that human–machine teams can cooperatively achieve a set of common goals. To understand the user's messages promptly, machine intelligence must substitute conventional data-driven machine learning approaches with CL between collaborative partners. The prerequisite for cooperation-oriented human–machine teaming is that the machine possesses a certain level of ToM: It can actively infer the user's intentions, desires, beliefs [103,104], and need for cooperation, thereby forming a human-centric and human-compatible process [106]. As the example in Section 2.3 suggests, the essence of establishing such cooperation lies in shared agency [107,108] or a common mind [3,109,110].

4.3.2. Prototype setting for generic human–robot teaming

We devised a human–robot collaboration system presented in the form of a cooperative game, in which a human user must work with a group of robotic scouts to complete certain tasks and optimize the group's gain [105]. In this game, the human user (the commander) and the robot scouts communicate over a restricted channel: Only the robot team interacts directly with the physical world; the commander does not directly access the physical world or directly control the behavior of the robot scouts. At the same time, only the commander has access to the true value function of the task (e.g., minimizing the total time); the robotics team must infer this value function through HRI. Such a setting realistically mimics real-world human–robot collaboration tasks, as many systems perform autonomously in hazardous environments under the supervision of human users.

To complete a game successfully, the robots must achieve bidirectional alignment by both "listening" and "speaking" wisely. First, the robots are expected to extract useful information from human feedbacks to infer the user's values and adjust their policies accordingly. Second, the robots are required to effectively explain what they have done and plan to do based on their current value inference, letting the user know whether or not the team shares the human values. Meanwhile, the commander is tasked to direct the robot scouts to reach the destination while maximizing the team's score. Therefore, the evaluation of a robot by the human is also a two-way process: The human user must infer the goal of the robot scouts, verify whether or not it aligns with the given value function of the task, and choose proper instructions to adjust the robots' goals if they are not aligned. Eventually, if the system performs well, the value function of the robot scouts should align well with the ground-truth value function provided only to the commander, and the commander should obtain high trust from the system. Fig. 12 illustrates the bidirectional value alignment process in the game [105]. There are three values in the interactive process:

- $U_{\mathscr{A}}$: the user's true value;
- $U_{\mathscr{A}\text{in}\mathscr{B}}$: the robot's estimation of the user's value, where, in this game, the scouts do not have their own value, so they will act according to $U_{\mathscr{A}\text{in}\mathscr{B}}$;
- $U_{\mathscr{B}\text{in}\mathscr{A}}$: the user's estimation of the robot's value, which is the ToM structure held by the user that is essential for feedback and trust formation.

Among these three values, two alignments take place:

- $U_{\mathscr{B}\text{in}\mathscr{A}} \rightarrow U_{\mathscr{A}}$: the robots learn the user's value from feedback.
- $U_{\mathscr{A}\text{in}\mathscr{B}} \rightarrow U_{\mathscr{B}\text{in}\mathscr{A}}$: the user learns the robots' value from explanation.

Eventually, the three values will converge to $U_{\mathscr{A}}$, at which moment the human–machine team will form mutual trust and effective collaboration.

This game design motivates spontaneous human–machine teaming and bidirectional reasoning, because both parties have crucial but private information at the beginning of the game. The robot scouts can obtain information about the map but lack access to the commander's value function. Since the value decides the mission goals, the robot scouts cannot make proper decisions reflecting the human user's intent on their own. In the meantime, the human user, despite knowing the task's value function that governs the decision-making process, cannot access the environment directly. By allowing constrained communication to fulfill human–machine collaboration, the robot scouts can make sporadic action proposals to the human user, and the human user provides binary accept/reject feedback, which the robot scouts then use to infer the correct value and adjust their actions accordingly.

Viewing this task from the perspective of the CL framework, we have the human user as the teacher A and the robot scouts team as the student B. The goal is for the robots' utility function to align with the user's and for the user to trust the robots—that is, for $u_{\text{A}} = u_{\text{B}}$ and $u_{\text{BinA}} = u_{\text{A}}$. As the robot team shares the environment information with the user in real time, we assume that their state beliefs are identical (as are their beliefs over state beliefs). This game does not entail model learning within the human–robot team, but the same algorithm for value alignment can easily be applied to model alignment with minor adjustments. m^t are the robots' proposals and accompanied explanations, while d^t are the user's feedback (acceptance or rejection) to the robot team. To achieve a fast alignment, the students need to know when and how to make proposals such that feedback from the teacher is the most informative to correct their value estimation. The feedback directly changes the belief of the robots. To obtain instructive feedback from the human teacher, the robots need to know what the human knows and believes, what she intends to do, and what is aligned and misaligned. It is only based on this shared agency and common mind that the robot scouts can provide proper explanations justifying their previous actions and current proposals, and influence the BoB of the human user appropriately.

4.3.3. Game setup

Our collaborative game has a minimal design and involves one human player as the commander and three robot scouts. The game's objective is to find a safe path in an unknown territory from

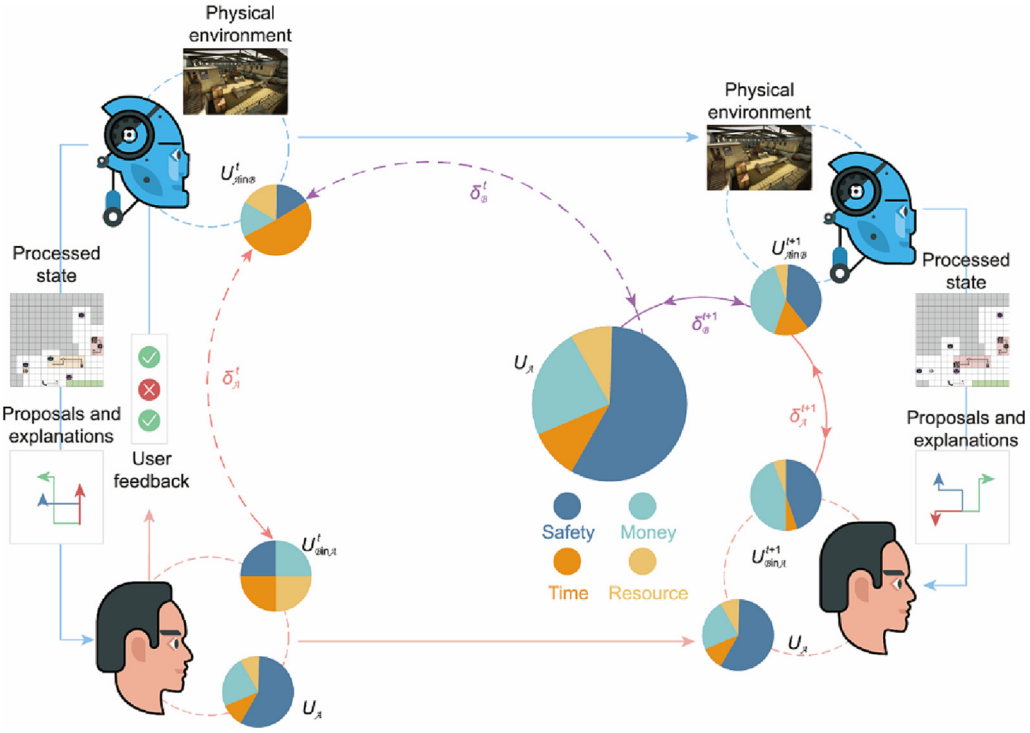

Fig. 12. Overview of bidirectional human–robot value alignment. Pie charts represent the values—that is, the importance of different goals in a collaboration task, such as simultaneously considering safety, gaining money, saving time, and reserving resources. t in the superscript represents the time step. $\mathscr{A}$ and $\mathscr{B}$ in the subscript represent "user" and "machine," respectively, indicating their resemblance to the teacher, A, and student, B, in the prior text. $U_{\mathscr{A}}$ is the user's true value, $U_{\mathscr{A}\text{in}\mathscr{B}}$ is the robot's estimation of the user's value, and $U_{\mathscr{B}\text{in}\mathscr{A}}$ is the user's estimation of the robot's current value. δ denotes the distance between values in the task value space. In every round of interaction, the machine first receives signals from the physical environment and processes its observations to form an abstract state of the environment. Next, the machine presents the processed map together with movement proposals and explanations to human users, who will provide feedback to the system accepting/rejecting the proposals according to human values and the current map state. Given the user's feedback, the machine then updates its estimation of human values and takes actions with respect to the new values. Cooperative human–robot communication with appropriate explanation aligns the team values in two directions by diminishing the distance between $U_{\mathscr{A}\text{in}\mathscr{B}}$ and $U_{\mathscr{A}}$, as well as $U_{\mathscr{B}\text{in}\mathscr{A}}$ and $U_{\mathscr{A}\text{in}\mathscr{B}}$, resulting in final convergence to the true value $U_{\mathscr{A}}$.

the base (located in the bottom right corner of the map) to the destination (located in the upper left corner of the map). The territory map is represented as a partially observed 20×20 tile board. In the map, every tile can either be empty or hold one of the various devices, which remains unobserved until a robot scout gets closer to it.

As they look for the safe path, the robot scouts can pursue a set of goals, including ① saving time for the path finding, ② scrutinizing suspicious devices on the map, ③ exploring tiles, and ④ collecting resources. The game performance of the human–robot team is measured by the accomplishment of these goals and their relative importance (weights), defined as the human user's value function, which is only known by the human user and not by the robot scouts. (See Fig. 13 for a snapshot of the game; Fig. 14 summarizes the human–machine interaction flow.)

4.3.4. Value alignment with CL

To estimate the human user's value during the communication process, we accommodate two levels of ToM into the computation model. The level-1 ToM models the cooperative assumption: A cooperative human user is more likely to accept proposals aligned with the correct value function than misaligned ones. The level-2 ToM further integrates the users' pedagogy into the model; that is, the user prefers the feedback that drives the robots' value closer to the true value over other feedbacks. We use a ToM one level deeper to delineate this pedagogical inclination, because it demands recursive modeling of the user's model of the robots. Incorporating both levels of ToM into our computational framework, we formulize the human interaction with functions parameterized by the value and develop a closed-form parameter learning algorithm [105].

It is notable that our human–machine teaming framework is a setting that is comparable to but different from IRL [111]. More specifically, IRL seeks to learn an underlying reward function that causes prerecorded expert demonstrations to satisfy certain optimality criteria in an offline passive learning setting. As we will summarize in Table 1 and specify in Appendix A, the standard IRL is a level-0 paradigm. In contrast, the robot scouts in our setting are designed to learn human value in an interactive way via sparse supervision from the user. Essentially, our design requires real-time and *in situ* inference of the human value as the human–robot collaboration proceeds—a unique property of human-centric learning schemes that fall into the category of level-2 paradigms. Furthermore, to consummate a collaboration, the robot and the user must have bidirectional alignment. That is, not only must the robot scouts quickly grasp the human user's intent, but they must also elucidate themselves to guarantee smooth communication with the commander throughout the entire game. In summary, the robots are tasked to drive bidirectional value alignment by actively inferring the human user's mental model, making proposals, and evaluating the user's feedback, all of which demand complex and recursive mind modeling of the collaborator.

5. Contributing to the foundations of learning

As discussed in Section 2, current frameworks in communication, applied math, and statistical machine learning are limited to special settings, and the derived performance bounds

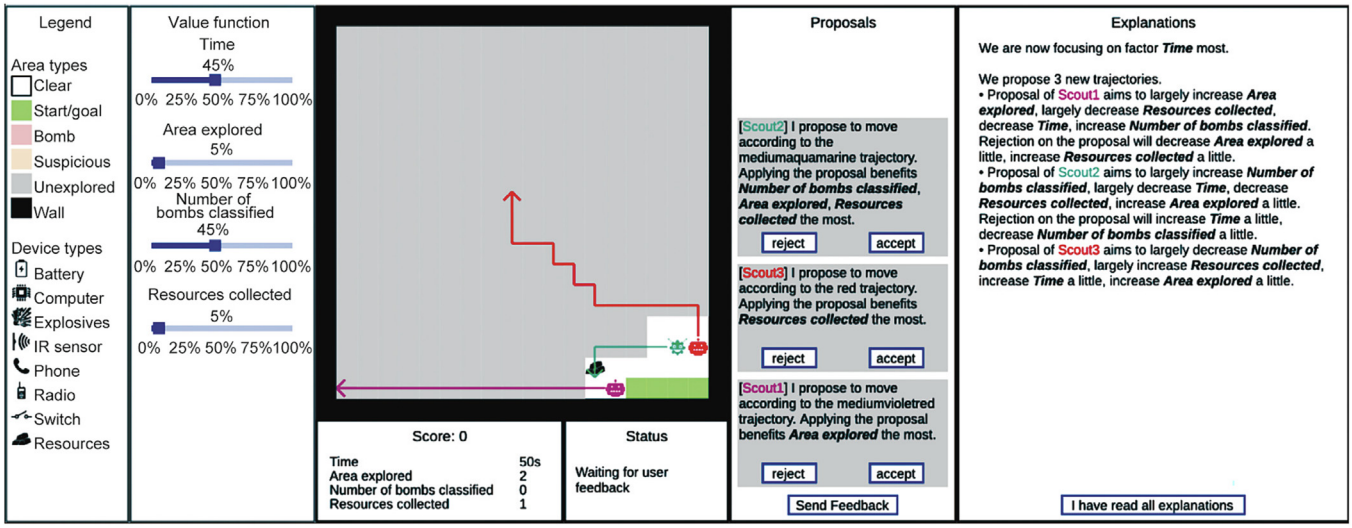

Fig. 13. User interface of the scout exploration game. From left to right, the legend panel explains the different types of tiles on the game map. The value function panel shows the current value function of the user's team. This value is given to the user at the beginning of the game but is unknown to the robot scouts and cannot be altered. The central map displays the current information on the game board. The score panel shows the user's current score and the individual fluent functions that contribute to the score. The overall score is calculated as the normalized, value function-weighted sum of the individual fluent function scores. The status panel updates the current status of the system to the user. The proposal panel shows the robot scouts' proposals in this round, and the user can accept/reject each. The explanation panel shows the explanations provided by the scouts.

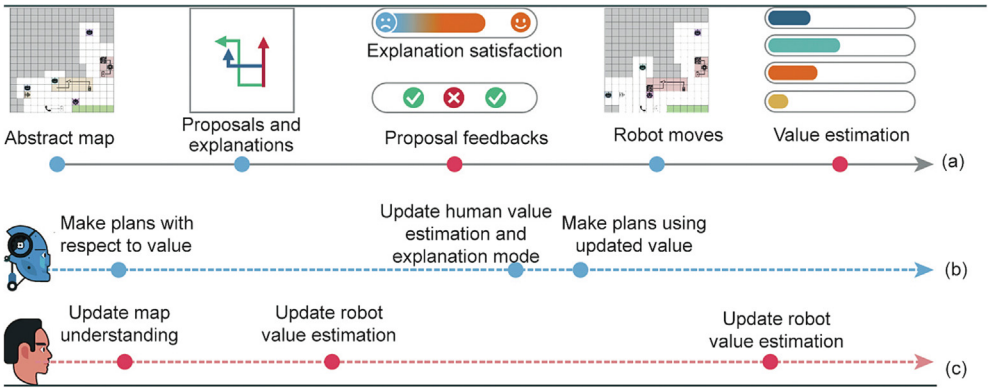

Fig. 14. Study design of the scout exploration game. (a) Timeline shows events in a single round of the game, beginning with the scouts receiving environment signals and ending with their next move. Users in different experimental groups receive proposals and explanations differently. The value estimation asks the users to infer the scouts' value at the current time. The answers to these questions are to probe the users' mental status and are only used in the analysis after the game completes. (b, c) Timelines depict the mental dynamics of the robots and the user, respectively.

[77,78,112–115] based on PAC learning and Vapnik–Chervonenkis (VC) dimensions are often overly pessimistic. Most concepts are "not learnable" in PAC learning settings [9], while human intelligence can learn from small examples for daily tasks. The main reason is that the current methods do not account for many important aspects of human communication and end up with less effective learning protocols.

In this section, we will elaborate on a few topics to formulate and develop the theoretical foundation to support CL. We first introduce a new representation of learning, a starting point of learning protocols going beyond the Shannon communication limit [8]. We then discuss the hierarchies of learning and put forth the halting problem of learning, on top of these hierarchies.

5.1. Learning representation introduced by CL

5.1.1. When Aumann meets Grice: From distributed knowledge to common knowledge

Realizing the insufficiency of Shannon's communication model and Valiant's PAC learning theory in modeling human learning, we seek another representation that is generic enough to accommodate both the scientist's and the student's way of learning. As discussed in Section 2.1, both types of learning can be interpreted as having information delivered from one mind to another. Hence, a generic representation should model both the "mind of departure" and the "mind of destination" of information delivery. In particular, we need a clear formulation for what is known and unknown in the learning process to model the transitions in the teacher's and the student's minds during learning. Fortunately, epistemic logic—that is, the logic of knowledge [116]—introduces a rigorous definition and mathematical representation of knowledge and beliefs. In the 1970s, Robert Aumann further expanded the existing logical analysis of knowledge reasoning to multiple agents and applied the concept of common knowledge to economics and game theories [117]. The notion of common knowledge, together with the later proposed concept of distributed knowledge [118], provides an ideal tool to represent agents' mental status before and after learning, and thus to model the whole learning process.

The framework for modeling knowledge is based on possible worlds [61]. The intuitive idea behind the possible-worlds model

is that, besides the true state of affairs, there are a number of other possible states of affairs or "worlds." Given its current information, an agent may not be able to tell which one of the possible worlds describes the actual state of affairs. An agent is then said to know a fact if it is true in all the possible worlds (given the agent's current information). Concretely, we can imagine a robot with a blurry camera. Visual signals received by the blurry camera are not clear enough for the robot to differentiate every possible world, so the robot needs to maintain a set of worlds that are all possible given the views it has received. The facts known by the robot must be true in all the worlds it considers possible.

When there are two agents reasoning about the world, the concepts of common knowledge and distributed knowledge become relevant.

- Common knowledge: Facts that both agents know, and both agents know that their partners know, and both agents know that their partners know that they know, and so on and so forth.
- Distributed knowledge: Facts that both agents will know if they fully combine their knowledge.

In other words, common knowledge is something that is known by both agents and that neither can deny, whereas distributed knowledge may not be known by any of the agents individually before they exchange their knowledge. Hence, distributed knowledge is always at least as precise as and usually more precise than common knowledge. The goal of learning is for the facts in the teacher and student's distributed knowledge to be delivered to their common knowledge. For example, suppose that there are two robots whose cameras are blurred in different ways. Learning takes place when they communicate and share their knowledge to prune both of their sets of possible worlds. When their common knowledge becomes identical to their distributed knowledge, the learning terminates, as both robots have no private information the other robots do not know.

We demonstrate an example of a learning process with logic knowledge representation in Fig. 15. Suppose that Alice and Bob have imperfect perception of the world (like the robot with a blurry camera). That is, they cannot observe ω; instead, they observe some projections as input.

$$I_{\mathrm{A}} = I_{\mathrm{A}}(\omega) = (I_{\mathrm{A},1}, ..., I_{\mathrm{A},8}) \tag{17}$$

$$I_{\mathrm{B}} = I_{\mathrm{B}}(\omega) = (I_{\mathrm{B},1}, ..., I_{\mathrm{B},8}) \tag{18}$$

where $I_{\mathrm{A}}(\omega)$ and $I_{\mathrm{B}}(\omega)$ project the world state, ω, into observation space, encoded by $I_{\mathrm{A},i}$ and $I_{\mathrm{B},i}$. Each input $I_{\mathrm{A},i}$, $I_{\mathrm{B},i} \in \{0, 1\}$ is binary and is $= 1$ if ω is on its right side and $= 0$ if ω is on the left. Their perception partitions are Π_{A} and Π_{B}, respectively.

$$\Pi_{\mathrm{A}}(\omega) = \{\omega' : I_{\mathrm{A}}(\omega') = I_{\mathrm{A}}(\omega)\} \tag{19}$$

$$\Pi_{\mathrm{B}}(\omega) = \{\omega' : I_{\mathrm{B}}(\omega') = I_{\mathrm{B}}(\omega)\} \tag{20}$$

That is, when the world is $\omega \in \Omega$, Alice cannot differentiate worlds in $\Pi_{\mathrm{A}}(\omega)$. That is, she knows the real world must be in $\Pi_{\mathrm{A}}(\omega)$, but she does not know which element of $\Pi_{\mathrm{A}}(\omega)$ it is. Similarly, Bob knows that one world in $\Pi_{\mathrm{B}}(\omega)$ must be true but is confused about the exact element in $\Pi_{\mathrm{B}}(\omega)$.

Then, with some preliminary definitions about partitions, we can define Alice and Bob's common and distributed knowledge. Given two partitions $\mathscr{P}$ and $\mathscr{P}'$ of a set S, then

- $\mathscr{P}$ is finer than $\mathscr{P}'$ if $\forall s \in S, \mathscr{P}(s) \subseteq \mathscr{P}'(s)$;
- $\mathscr{P}$ is coarser than $\mathscr{P}'$ if $\forall s \in S, \mathscr{P}'(s) \subseteq \mathscr{P}(s)$.

Intuitively, if partition $\mathscr{P}$ is finer than partition $\mathscr{P}'$, then the information sets given by $\mathscr{P}$ give at least as much information as the information sets given by $\mathscr{P}'$ (since considering fewer states possible corresponds to having more information). The meet of partitions $\mathscr{P}$ and $\mathscr{P}'$, denoted $\mathscr{P} \cap \mathscr{P}'$, is the finest partition that is coarser than $\mathscr{P}$ and $\mathscr{P}'$; the join of $\mathscr{P}$ and $\mathscr{P}'$, denoted as $\mathscr{P} \cup \mathscr{P}'$, is the coarsest partition finer than $\mathscr{P}$ and $\mathscr{P}'$ (see Chapter 3 in Ref. [61]). Next, we can make use of the meet and the join to give nice characterizations of common knowledge and distributed knowledge.

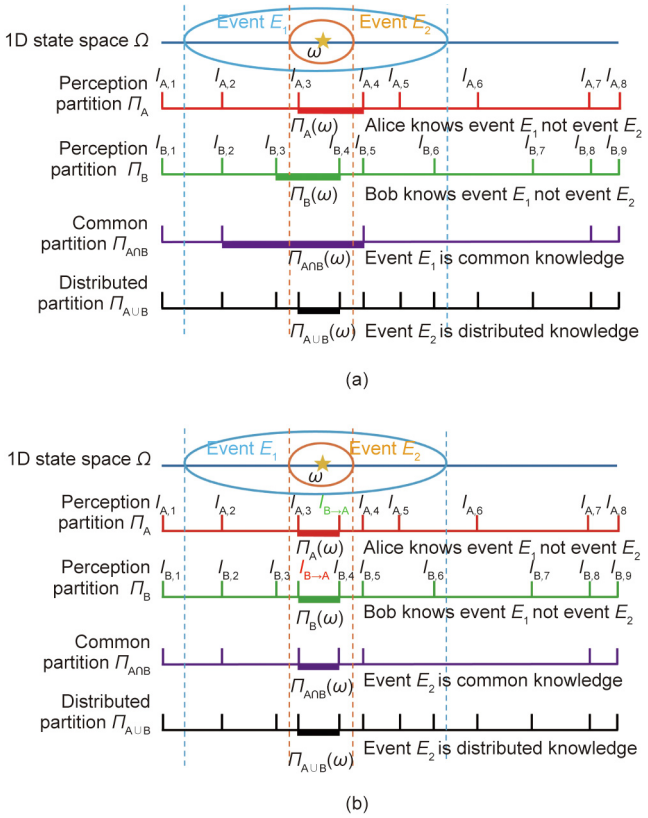

Fig. 15. Common and distributed knowledge for inferring a state ω (star) in the one-dimensional (1D) space between Alice and Bob. Every segment represents a cell in the partition. States that have fallen into the same segment cannot be differentiated. (a) Knowledge representation before communication; (b) knowledge representation after communication. Alice and Bob's perception partition becomes finer because the partners' messages enable further differentiation of worlds. Π_{A} and Π_{B} represent perception partitions; $\Pi_{\mathrm{A}\cap\mathrm{B}}$ and $\Pi_{\mathrm{A}\cup\mathrm{B}}$ is the meet and the join of the two partitions Π_{A} and Π_{B}, respectively. $I_{\mathrm{A},1}, ..., I_{\mathrm{A},8}$ and $I_{\mathrm{B},1}, ..., I_{\mathrm{B},8}$ are projections of observations.

As shown in Fig. 15(a), before they communicate, the meet of the two partitions, $\Pi_{\mathrm{A}\cap\mathrm{B}}$, forms the common partition of Alice and Bob, and the join of the two partitions, $\Pi_{\mathrm{A}\cup\mathrm{B}}$, forms the distributed partition of them. That the event E_1 happens is the common knowledge, because

$$\Pi_{\mathrm{A}\cap\mathrm{B}}(\omega) \subset E_1 \tag{21}$$

That is, because the event E_1 happens in all worlds of $\Pi_{\mathrm{A}\cap\mathrm{B}}(\omega)$, both Alice and Bob know that it happens, although they do not know the exact world. Moreover, neither Alice nor Bob can deny that they do not know E_1, because $\Pi_{\mathrm{A}\cap\mathrm{B}}(\omega)$ contains all possible confused worlds for both Alice and Bob[†], making E_1 their common knowledge. On the contrary, neither Alice nor Bob knows that the event E_2 happens, because

[†] Intuitively, since Alice knows that $\omega \in \Pi_{\mathrm{A}}(\omega)$, she knows that Bob must know that $\omega \in [I_{\mathrm{B},3}, I_{\mathrm{B},5}]$. Likewise, Bob knows that Alice knows that $\omega \in [I_{\mathrm{A},2}, I_{\mathrm{A},4}]$. Together, the mutual knowledge forms $\Pi_{\mathrm{A}\cap\mathrm{B}}(\omega)$. Rigorously, the definition of common knowledge triggers infinite recursions. In the case of Fig. 15, the recursion converges after one round.

$$\Pi_{\mathrm{A}}(\omega) \not\subset E_2 \wedge \Pi_{\mathrm{B}}(\omega) \not\subset E_2 \tag{22}$$

However, by combining their knowledge, E_2 is also commonly knowable, because

$$\Pi_{\mathrm{A}\cup\mathrm{B}}(\omega) \subset E_2 \tag{23}$$

In Fig. 15(b), we show that, by sharing one of their partition boundaries with their partner, Alice and Bob can transit the event E_2 from their distributed knowledge to their common knowledge. That is,

$$m_{\mathrm{A}\to\mathrm{B}} = I_{\mathrm{A},3}(\omega) \quad \text{and} \quad m_{\mathrm{B}\to\mathrm{A}} = I_{\mathrm{B},4}(\omega) \tag{24}$$

Then, through one round of messaging, E_2 becomes common knowledge. The perceived cell is compressed:

$$\Pi_{\mathrm{A}}(\omega) = \{\omega' : I_{\mathrm{A}}(\omega') = I_{\mathrm{A}}(\omega) \wedge I_{\mathrm{A}}(\omega') = m_{\mathrm{B}\to\mathrm{A}}\} \tag{25}$$

$$\Pi_{\mathrm{B}}(\omega) = \{\omega' : I_{\mathrm{B}}(\omega') = I_{\mathrm{B}}(\omega) \wedge I_{\mathrm{B}}(\omega') = m_{\mathrm{A}\to\mathrm{B}}\} \tag{26}$$

Here, we slightly abuse the notation and use the equal sign to indicate a perception that is consistent with the received message. To summarize, using the framework of logical knowledge analysis, learning is modeled as the agents communicating and sharing their individual knowledge so that their partition of the worlds is refined and information is delivered from distributed knowledge to common knowledge.

5.1.2. Beyond Shannon's limits: A better learning protocol enabled by CL

The notion of common and distributed knowledge provides a formal representation of learning. Nevertheless, the knowledge representation alone still falls short of modeling the cooperation in human pedagogy. More specifically, information delivery from distributed knowledge to common knowledge answers the question of what learning is, but it does not approach how the teacher and the student should send out and comprehend messages—that is, what an efficient learning protocol is. To answer the question of how to learn efficiently, we must arm Aumann's knowledge representation with pragmatics.

As we mentioned in Section 4.1.2, pragmatics is the branch of linguistics that studies how the context of language using contributes to the meanings [10,36,119]. Recall the example illustrated in Figs. 5, 9, and 10. In the context of pedagogy, where the teacher and the student form a collaborative group, not only are the literal meanings of the teacher's messages considered, but her actions of choosing certain messages over others also facilitate the student's learning. The student makes exquisitely sensitive inferences about what the utterance means, given their knowledge of the situation, the context, and the teacher [83]. A famous concretization is the Gricean maxim of quantity [10] or the scalar implicature: When people say, "I like drinking warm coffee," although the lexical meaning of "warm" is semantically close to "hot," they mean "not hot;" otherwise, the people would have said "hot" directly [120,121]. This simple phenomenon entails two fundamental characteristics of human communication: the collaborative common ground between the listener and speaker, and recursive ToM modeling. By embedding learning in a communication framework, CL can satisfy both of these conditions that are impossible to meet using unilateral machine learning paradigms.

In Fig. 16, we give an example demonstrating the advantage of integrating pragmatic reasoning into the learning protocol. Let state ω be an image, and let Alice and Bob have N_{A} and N_{B} neurons as their observations:

$$I_{\mathrm{A}}(\omega) = \left(h_{\mathrm{A}}^{1}(\omega), \ldots, h_{\mathrm{A}}^{N_{\mathrm{A}}}(\omega)\right) \tag{27}$$

$$I_{\mathrm{B}}(\omega) = \left(h_{\mathrm{B}}^{1}(\omega), \ldots, h_{\mathrm{B}}^{N_{\mathrm{B}}}(\omega)\right) \tag{28}$$

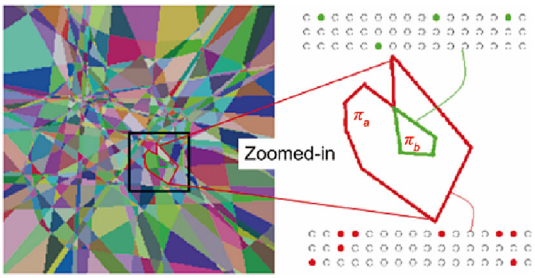

Fig. 16. An example of partition in 2D space and inference of a state using the pragmatic protocol. Left: The partition of the 2D space. States within the same patch trigger identical neural responses. Right: neurons fired for π_a (red) and π_b (green), cells marking two sets of possible states.

Here h denotes neurons, which can be indicators or ReLU projections of the image. That is,

$$h(\omega) = 1(\langle \omega, \lambda \rangle \geq 0) \quad \text{or} \quad \max(0, \langle \omega, \lambda \rangle) \tag{29}$$

where λ is the weight of the neuron. As shown in Fig. 16, π_a is surrounded by eight red neurons, while π_b is bounded by four green neurons. Suppose that Alice knows an event $\omega \in \pi_a$ and tells Bob via the following message

$$m_{\mathrm{A}\to\mathrm{B}} = \left(h_{\mathrm{A},a_1}, \ldots, h_{\mathrm{A},a_8}\right) \tag{30}$$

Bob will then refine his perception from $\Pi_{\mathrm{B}}(\omega)$ to π_a, achieving an information gain of

$$\mathrm{IG}_{\mathrm{Shannon}} = \log_2 \frac{|\Pi_{\mathrm{B}(\omega)}|}{|\pi_a|} \tag{31}$$

as defined in Eq. (1). That is, Bob's belief of the possible worlds narrows down from $\Pi_{\mathrm{B}}(\omega)$ to π_a. Interestingly, if Bob has ToM and integrates pragmatics into the learning protocol, he will read between the lines: Since Alice could but did not send a shorter message using the four green neurons, she must imply not π_b but π_a. Therefore, Bob can further refine his belief and achieve an information gain of

$$\mathrm{IG}_{\mathrm{CL}} = \log_2 \frac{|\Pi_{\mathrm{B}}(\omega)|}{|\pi_a / \pi_b|} \tag{32}$$

where π_a/π_b means areas in π_a but not in π_b.

Proposition. The pragmatic protocol is more effective than Shannon's communication protocol, as Bob gains more information than Shannon's information measurement, since we have

$$\mathrm{IG}_{\mathrm{CL}} > \mathrm{IG}_{\mathrm{Shannon}} \tag{33}$$

The pragmatic protocol goes beyond Shannon's information limit by integrating ToM. The extra information gain is brought by reflecting the minds of the other agents: Alice selects messages after deliberating what Bob knows, and Bob reasons why Alice sent this message instead of other plausible messages.

5.2. The halting problem of learning

After studying the convergence from distributed knowledge to common knowledge, we raise the "halting problem of learning," by analogy to the halting problem of computing [122]. That is, under what conditions does the learning process terminate at various equilibria, which define the fundamental limits of learning. Just like the pedagogy and learning in our everyday lives, CL proceeds iteratively. For iterative learning, the problem of halting is ubiquitous. Thus far in this article, we have not dived into this problem. In Algorithm 1, the algorithms terminate when a fixed number of steps is reached. However, deciding the proper rounds

of interaction beforehand can be challenging, if not impractical. Thus, we need some criteria to monitor the learning and terminate the process when learning arrives at its limits.

To approach the halting problem of learning, we must know the underlying driving force of learning—namely, what do the teacher and the student seek to agree on when they communicate with each other? Here, we recognize the three granularity of learning:

(1) **Message level:** an understanding of messages in a single round of communication;

(2) **Task level:** mental alignment between the teacher and the student about a specific task, involving multiple rounds of communication;

(3) **Group level:** an understanding of the partner's characterization, reused across different tasks.

Every level has a distinctive goal and is controlled by a loop in CL, as depicted in Fig. 17. Within each loop, the teacher and the student aim to achieve an equilibrium. In the next section, we introduce each level in detail. As one purpose of CL is to facilitate the development of new learning paradigms, we also include opening questions along with the introduction to encourage future explorations in related topics.

5.2.1. Three hierarchies in learning

5.2.1.1. Message level. The message level indicates the interpretation of each message between the teacher and the student. As messages are the building blocks of a communication process, being able to fully comprehend the speaker is the prerequisite of an effective learning process. At this level, messages communicate about a state, a cell, or an event (set) in the state space in order to achieve a common ground and common belief. Although we only discuss a few types of messages such as labeled data (Section 3.4), projection on cells as partitions (Section 5.1.1), and linear neurons (Section 5.1.2), the message space of CL can be extended to nodes in a parse graph or logic predicates that correspond to the compositions of atomic cells or events. The reflection loop in CL is in charge of the equilibrium in the message level. The reflection processes involve loops, because the agents bear ToM and conduct recursive mutual reasoning. In particular, the teacher considers what the student needs to know, and the student thinks about why the teacher sends one specific message instead of others, and so forth. Hence, the agents form reflection loops entailing their egocentric belief b_θ/b_ω and their ToM belief $b_s/b_{\tilde{s}}$, as shown in Fig. 17.

At this level, ideal agents should be capable of capturing both the literal and pragmatic meaning of the messages. To date, most works assume that the teacher and the student have a common ground about the literal meanings of all the messages. In other words, they speak the same language; or, at least, each has a dictionary for the language the other is speaking. One relaxation of this assumption is to take away their dictionaries and see if they can still develop a valid codebook that they share as the common literal meanings of the messages. Namely, the teacher needs to learn ϕ, ψ and the student needs to learn ξ, ζ, and π from scratch as a group or individually with the partner being fixed. Another challenge at this level is the derivation of pragmatic meanings, which often requires counterfactual reasoning. That is, to understand the speaker, the listener needs to consider not only the explicit selected message but also the implicit unselected messages. When the message space is large, such as the English language, such counterfactual reasoning often becomes intractable and requires an efficient method to be fully interpreted.

5.2.1.2. Task level. Task-level learning involves transmitting and understanding a sequence of messages to obtain the convergence of the teacher's and the student's mental components. The learning loop in CL aims to find the equilibrium in this level—that is, for the student's utility, model, and policy to be close enough to the teacher's counterparts. To evaluate the convergence, agents need to (either directly or indirectly) measure the distance between their minds. Direct measurements rely on the representation of the mental components, such as the Kullback–Leibler (KL) divergence for beliefs with a closed form. Indirect measurements can be applied by comparing the task performance between the teacher and the student, such as a sufficient statistics difference between their motion trajectories. We denote this evaluation metric as $L\left(b_\theta, b_\omega^T\right)$ in Fig. 17.

Since teaching often requires a sequence of messages, task-level learning accommodates planning for sequential pedagogy, in which the communication takes multiple steps. As the planning complexity is exponential with respect to the number of steps,

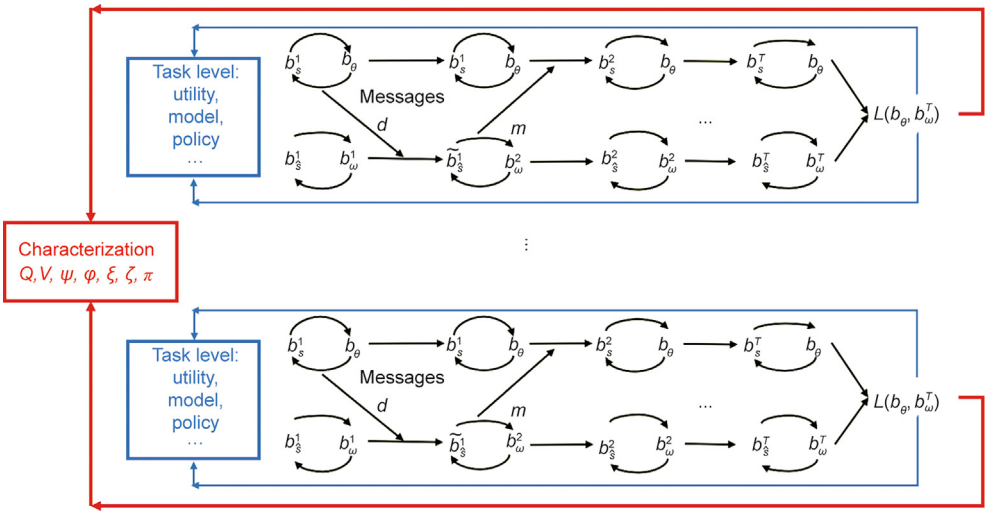

Fig. 17. CL includes three nested loops: ① The reflection loop in black for the deliberation of inferential messages to achieve common ground; ② the learning loop in blue to achieve a common model, utility, policy, and so forth; and ③ the characterization loop in red to achieve a better group understanding of each agent's characterization, such as the value and belief-update functions defined in Section 3.3. The norm of communication forms when group members' estimations of their partners' characterizations stabilize. b_θ and b_ω are beliefs of the model; b_s, $b_{\tilde{s}}$, and $\tilde{b}_{\tilde{s}}$ are beliefs of the learning state; m and d are messages from the student and the teacher, respectively; L denotes the loss function and superscripts represent time stamp. All notations follow the definition in Section 3.3.

and the message space is usually large or dynamic, most existing teachers use heuristics or short-horizon planning. Nevertheless, even for this inadequate planning, the student is treated as a naive learner who does not model the teacher (a level-1 paradigm). Developing a planning algorithm for teachers in level-2 paradigms remains an active research question. It requires the teacher and the student to form a group-level norm of communication, which falls under the scope of group-level learning. Another potential improvement of task-level learning is to generalize message-level pragmatic reasoning to the task level; or, in other words, to endow the student with more information from the unselected teaching sequences in addition to the selected one.

5.2.1.3. Group level. Finally, the group level entails the characterization of the teacher and the student. To conduct effective CL, the value and belief-update functions of the teacher and the student must work in accordance. In most prior learning paradigms, the characterizations of the agents are predefined by heuristics. In more generic and realistic settings, the teacher and the student need to learn proper characterization in order to collaborate. The characterization loop in CL seeks to tackle this problem. Just as human teams demand sufficient collaboration experience to acquire tacit cooperation, it usually requires multiple learning tasks for the agents to learn appropriate characterization. As shown in Fig. 17, the outcomes of multiple learning tasks will shape the agents' characterizations. The referential game in Section 4.1 follows this process. Afterward, norms of communication and a learning protocol will be established between the teacher and the student.

In all the works surveyed in this article, one teacher is exclusively designed for one or one type of student. A standard characterization loop also only sets up the learning protocol for one specific group. A more ambitious setting is to have a teacher that can adapt to different kinds of students—even those she never encountered during her training. Suppose that we can characterize the student's personality, such as his IQ and memory, which correspond to his mental dynamics; then, a versatile teacher should be able to identify her student's characterization and customize her pedagogy accordingly. To be more specific, the teacher (student) can parameterize her (his) dynamic functions, ψ, $\phi\,(\xi,\,\zeta,\,\pi)$, and model a distribution of the parameters. Such a setting is similar to the *ad hoc* teamwork [123] and multitask/meta-RL in MDPs with different dynamics [124], but it has not been thoroughly studied in the context of pedagogical machine learning.

5.2.2. Halting criteria

Knowing the three hierarchies of learning, we can discuss the appropriate halting criteria of learning. In every level, the CL loop terminates under different conditions. In the message level, a plausible halting criterion is fully understanding both the literal and pragmatic meaning of each message. The literal meaning is usually straightforward when agents use mutually recognizable message spaces. The pragmatic meaning depends on the context, such as the situated state, the communication history, and so on. As pragmatic understanding requires ToM, the reflection loop usually stops when the recursive reasoning converges or when the cognitive burden of going to the next recursion layer exceeds the cost of sending an explicit query message to clarify.

In the task level, the ideal halting occurs when the six minds converge to one and are validated by the oracle (God's mind).

Recall the high-level mind notation we used in Fig. 1. That is,

$$P_t = Q_t = \widehat{P}_t = \widehat{Q}_t = C_t = G \tag{34}$$

This is the strictest halting condition. In many cases, some alternative halting conditions can be defined. For example, let $\mathscr{D}(X, Y)$ denote the distance between two minds X and Y, such as the KL divergence, total variance distance, earth mover's distance, and so forth. Then, we can name a few halting conditions:

- $\mathscr{D}\left(P_t, \widehat{Q}_t\right) \leq \epsilon$: The teacher thinks that the student already knows what she knows and stops teaching.
- $\mathscr{D}(Q_{t-1}, Q_t) \leq \epsilon$: The student becomes complacent and thinks that he cannot acquire new knowledge. Thus, he stops learning.
- $\mathscr{D}\left(\widehat{Q}_{t-1}, \widehat{Q}_t\right) \leq \epsilon$: The teacher thinks that the student cannot make further promising progress and stops teaching.
- $\mathscr{D}(C_t, \phi) \leq \epsilon$: The teacher and the student find it difficult to reach common ground and terminate communication.
- $\mathscr{D}(Q_t, G) \leq \epsilon$: The student achieves satisfactory performance in the real world and stops learning. Notice that, in some cases, G may not be directly accessible, so certain surrogate functions may be needed; for example, it may be necessary to let the student finish some unseen tasks as tests.

The above conditions are by no means comprehensive; more can be proposed to meet distinctive needs and circumstances. In addition, conditions more complicated than a single distance function can also be defined, such as comparing the gain of minimizing the divergence between the student and the teacher against the cost of transmitting teaching messages.

In the group level, the teacher and the student can stop interacting with each other after they know their partners' characteristics, that is, dynamic functions, state, model, value spaces, and so forth. Accomplishing this usually requires collaboration over several pedagogy tasks. The assessment of the convergence needs the teacher and the student to communicate given unseen learning objectives. It is only when they can cooperate effectively in diverse tasks that they can terminate the characterization loop and halt group-level learning. Suppose that we also want a flexible teacher who can adapt to various students quickly. Then, the halting condition will be for the teacher to successfully conduct CL with multiple students, each covering multiple learning objectives, resembling the evaluation used in *ad hoc* teaming [125].

6. Conclusions

In this article, we study a CL formalism, which inspects learning from a communication perspective. We review existing learning paradigms such as passive learning, active learning, and algorithmic teaching and examine their limitations. The new formalism has the potential to overcome these limitations and integrate prior machine learning algorithms into a unified framework. With concrete usage examples, we demonstrate how efficient learning protocols can emerge from CL and verify the formalism's suitability for complicated HRI tasks and generic human–machine collaboration. Moreover, the CL formalism makes two contributions to the foundation of learning: First, it introduces new representations of learning and puts forth protocols beyond Shannon's communication limits. Second, it sheds light on the universal halting problem of learning by teasing out three hierarchies in learning and defining possible halting criteria. Overall, we see CL as providing conceptual clarity for existing and future methods of machine learning from the perspective of mutual reasoning between the teacher and the student, and as a fruitful base for future work on more advanced learning paradigms.

Acknowledgments

The works in China from the authors reported herein are supported by a National Key Research and Development Program of China (2022ZD0114900), and the works at University of California,

Los Angeles were supported by Multidisciplinary Research Program of the University Research Initiative Office of Naval Research (MURI ONR; N00014-16-1-2007) and Defense Advanced Research Projects Agency Explainable Artificial Intelligence DARPA XAI (N66001-17-2-4029).

Compliance with ethics guidelines

Luyao Yuan and Song-Chun Zhu declare that they have no conflict of interest or financial conflicts to disclose.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.eng.2022.10.017.

References

[1] Zhu Y, Gao T, Fan L, Huang S, Edmonds M, Liu H, et al. Dark, beyond deep: a paradigm shift to cognitive AI with humanlike common sense. Engineering 2020;6(3):310–45.
[2] Shulman LS. Knowledge and teaching: foundations of the new reform. Harv Educ Rev 1987;57(1):1–23.
[3] Tomasello M. Origins of human communication. Cambridge: MIT Press; 2010.
[4] Holyoak KJ, Thagard P. Mental leaps: analogy in creative thought. Cambridge: MIT Press; 1995.
[5] Lake BM, Salakhutdinov R, Tenenbaum JB. Human-level concept learning through probabilistic program induction. Science 2015;350(6266):1332–8.
[6] Premack D, Woodruff G. Does the chimpanzee have a theory of mind? Behav Brain Sci 1978;1(4):515–26.
[7] Clark HH. Using language. New York City: Cambridge University Press; 1996.
[8] Shannon CE. A mathematical theory of communication. Bell Syst Tech J 1948;27(3):379–423.
[9] Valiant LG. A theory of the learnable. Commun ACM 1984;27(11):1134–42.
[10] Grice HP. Logic and conversation. In: Cole P, Morgan J, editors. Syntax and semantics: speech acts. New York City: Academic Press; 1975.
[11] Levinson SC. Presumptive meanings: the theory of generalized conversational implicature. Cambridge: MIT Press; 2000.
[12] Goodman ND, Stuhlmüller A. Knowledge and implicature: modeling language understanding as social cognition. Top Cogn Sci 2013;5:173–84.
[13] Eaves BS, Schweinhart Jr AM, Shafto P. Tractable Bayesian teaching. In: Jones M, editor. Big data in cognitive science. New York City: Psychology Press; 2015.
[14] Eaves Jr BS, Feldman NH, Griffiths TL, Shafto P. Infant-directed speech is consistent with teaching. Psychol Rev 2016;123(6):758–71.
[15] Ho MK, Littman ML, MacGlashan J, Cushman F, Austerweil JL. Showing versus doing: teaching by demonstration. In: Lee D, Sugiyama M, Luxburg U, Guyon I, Garnett R, editors. Advances in neural information processing systems. Barcelona: Curran Associates, Inc.; 2016.
[16] Samuel AL. Some studies in machine learning using the game of checkers. IBM J Res Dev 1959;3(3):210–29.
[17] Bishop CM. Pattern recognition and machine learning. New York City: Springer; 2006.
[18] Shalev-Shwartz S, Ben-David S. Understanding machine learning: from theory to algorithms. New York City: Cambridge University Press; 2014.
[19] Deng J, Dong W, Socher R, Li LJ, Li K, Li FF. Imagenet: a large-scale hierarchical image database. In: Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition; 2009 Jun 20–25; Miami, FL, USA; 2009.
[20] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Commun ACM 2017;60(6):84–90.
[21] Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2014 Jun 23–28; Columbus, OH, USA; 2014.
[22] Girshick R. Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision; 2015 Dec 11–18; Santiago, Chile; 2015.
[23] He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision; 2017 Oct 22–29; Venice, Italy; 2017.
[24] Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. 2018. arXiv:1810.04805.
[25] Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D, et al. Playing Atari with deep reinforcement learning. 2013. arXiv:1312.5602.
[26] Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, et al. Mastering the game of Go with deep neural networks and tree search. Nature 2016;529(7587):484–9.
[27] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 27–30; Las Vegas, NV, USA; 2016.
[28] Angluin D. Queries and concept learning. Mach Learn 1988;2(4):319–42.
[29] Settles B. Active learning literature survey. Technical report. Madison: University of Wisconsin-Madison; 2010.
[30] Argall BD, Chernova S, Veloso M, Browning B. A survey of robot learning from demonstration. Robot Auton Syst 2009;57(5):469–83.
[31] Shafto P, Goodman ND, Griffiths TL. A rational account of pedagogical reasoning: teaching by, and learning from, examples. Cogn Psychol 2014;71:55–89.
[32] Milli S, Abbeel P, Mordatch I. Interpretable and pedagogical examples. 2017. arXiv:1711.00694.
[33] Yang SCH, Yu Y, Givchi A, Wang P, Vong WK, Shafto P. Optimal cooperative inference. In: Proceedings of the 21st International Conference on Artificial Intelligence and Statistics; 2018 Apr 9–11; Lanzarote, Spain; 2018.
[34] Chen Y, Aodha OM, Su S, Perona P, Yue Y. Near-optimal machine teaching via explanatory teaching sets. In: Proceedings of the 21st International Conference on Artificial Intelligence and Statistics; 2018 Apr 9–11; Lanzarote, Spain; 2018.
[35] Chen Y, Singla A, Aodha OM, Perona P, Yue Y. Understanding the role of adaptivity in machine teaching: the case of version space learners. 2018. arXiv:1802.05190.
[36] Hadfield-Menell D, Russell SJ, Abbeel P, Dragan A. Cooperative inverse reinforcement learning. In: Lee D, Sugiyama M, Luxburg U, Guyon I, Garnett R, editors. Advances in neural information processing systems. Barcelona: Curran Associates, Inc.; 2016.
[37] Ho MK, Littman ML, Cushman F, Austerweil JL. Effectively learning from pedagogical demonstrations. In: Proceedings of the Annual Conference of the Cognitive Science Society; 2018 Jul 25–28; Madison, WI, USA; 2018.
[38] Cakmak M, Lopes M. Algorithmic and human teaching of sequential decision tasks. In: Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence; 2012 Jul 22–26; Toronto, ON, Canada; 2012.
[39] Zhu X. Machine teaching for Bayesian learners in the exponential family. In: Proceedings of the 27th International Conference on Neural Information Processing Systems; 2013 Dec 9–12; Lake Tahoe, NV, USA; 2013.
[40] Zhu X. Machine teaching: an inverse problem to machine learning and an approach toward optimal education. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence; 2015 Jan 25–30; Austin, TX, USA; 2015.
[41] Liu W, Dai B, Humayun A, Tay C, Yu C, Smith LB, et al. Iterative machine teaching. In: Proceedings of the 34th International Conference on Machine Learning; 2017 Aug 6–11; Sydney, NSW, Australia; 2017.
[42] Fan Y, Tian F, Qin T, Li XY, Liu TY. Learning to teach. In: Proceedings of the 6th International Conference on Learning Representations; 2018 Apr 30–May 3; Vancouver, BC, Canada; 2018.
[43] Jiang L, Zhou Z, Leung T, Li LJ, Li FF. MentorNet: learning data-driven curriculum for very deep neural networks on corrupted labels. In: Proceedings of the 35th International Conference on Machine Learning; 2018 Jul 10–15; Stockholm, Sweden; 2018.
[44] Han B, Yao Q, Yu X, Niu G, Xu M, Hu W, et al. Co-teaching: robust training of deep neural networks with extremely noisy labels. In: Proceedings of the 32th Conference on Neural Information Processing Systems; 2018 Dec 3–8; Montreal, QC, Canada; 2018.
[45] Wang P, Wang J, Paranamana P, Shafto P. A mathematical theory of cooperative communication. In: Proceedings of the 34th Conference on Neural Information Processing Systems; 2020 Dec 6–12; Vancouver, BC, Canada; 2020.
[46] Gweon H, Tenenbaum JB, Schulz LE. Infants consider both the sample and the sampling process in inductive generalization. Proc Natl Acad Sci USA 2010;107(20):9066–71.
[47] Csibra G, Gergely G. Social learning and social cognition: the case for pedagogy. In: Munakata Y, Johnson MH, editors. Processes of change in brain and cognitive development—attention and performance XXI. Oxford: Oxford University Press; 2006.
[48] Csibra G, Gergely G. Natural pedagogy. Trends Cogn Sci 2009;13(4):148–53.
[49] Xu F, Denison S. Statistical inference and sensitivity to sampling in 11-month-old infants. Cognition 2009;112(1):97–104.
[50] Xu F, Tenenbaum JB. Sensitivity to sampling in Bayesian word learning. Dev Sci 2007;10(3):288–97.
[51] Gweon H, Shafto P, Schulz L. Development of children's sensitivity to overinformativeness in learning and teaching. Dev Psychol 2018;54(11):2113–25.
[52] Sperber D, Wilson D. Relevance: communication and cognition. Oxford: Blackwell; 1986.
[53] Peltola T, Çelikok MM, Daee P, Kaski S. Machine teaching of active sequential learners. In: Proceedings of the 33th Conference on Neural Information Processing Systems; 2019 Dec 8–14; Vancouver, BC, Canada; 2019.
[54] Shafto P, Goodman N. Teaching games: statistical sampling assumptions for learning in pedagogical situations. In: Proceedings of the 30th Annual Conference of the Cognitive Science Society; 2008 Jul 23–26; Washiton, DC, USA; 2008.
[55] Wang J, Wang P, Shafto P. Sequential cooperative Bayesian inference. In: Proceedings of the 37th International Conference on Machine Learning; 2020 Jul 13–18; Vienna, Austria; 2020.
[56] Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: data mining, inference, and prediction. New York City: Springer; 2009.
[57] Vapnik V. The nature of statistical learning theory. New York City: Springer; 1999.

[58] Rivest RL. Cryptography and machine learning. In: Proceedings of the International Conference on the Theory and Applications of Cryptology: Advances in Cryptology; 1991 Nov 11–14; Fujiyoshida, Japan; 1991.
[59] Zilles S, Lange S, Holte R, Zinkevich MA. Models of cooperative teaching and learning. J Mach Learn Res 2011;12:349–84.
[60] Weaver W. Recent contributions to the mathematical theory of communication. ETC Rev Gen Semant 1953;10(4):261–81.
[61] Fagin R, Halpern JY, Moses Y, Vardi MY. Reasoning about knowledge. Cambridge: MIT Press; 2003.
[62] Doshi P, Gmytrasiewicz PJ. Monte Carlo sampling methods for approximating interactive POMDPs. J Artif Intell Res 2009;34:297–337.
[63] Albrecht SV, Stone P. Autonomous agents modelling other agents: a comprehensive survey and open problems. Artif Intell 2018;258:66–95.
[64] Foerster J, Chen RY, Al-Shedivat M, Whiteson S, Abbeel P, Mordatch I. Learning with opponent-learning awareness. In: Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems; 2018 Jul 10–15; Stockholm, Sweden; 2018.
[65] De Weerd H, Verbrugge R, Verheij B. Theory of mind in the Mod game: an agent-based model of strategic reasoning. In: Proceedings of the European Conference on Social Intelligence; 2014 Nov 3–5; Barcelona, Spain; 2014.
[66] De Weerd H, Verbrugge R, Verheij B. Higher-order theory of mind in the Tacit Communication Game. Biol Inspired Cogn Archit 2015;11:10–21.
[67] Zhu SC, Mumford D. A stochastic grammar of images. Found Trends Comput Graph Vis 2007;2(4):259–362.
[68] Qi S, Zhu Y, Huang S, Jiang C, Zhu SC. Human-centric indoor scene synthesis using stochastic grammar. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR); 2018 Jun 18–22; Salt Lake City, UT, USA; 2018.
[69] Liu C, Chai JY, Shukla N, Zhu SC. Task learning through visual demonstration and situated dialogue. In: Workshops at the Thirtieth AAAI Conference on Artificial Intelligence; 2016 Feb 12–17; Phoenix, AZ, USA; 2016.
[70] Liu C, Yang S, Saba-Sadiya S, Shukla N, He Y, Zhu SC, et al. Jointly learning grounded task structures from language instruction and visual demonstration. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing; 2016 Nov 1–5; Austin, TX, USA; 2016.
[71] Shukla N, He Y, Chen F, Zhu SC. Learning human utility from video demonstrations for deductive planning in robotics. In: Proceedings of Conference on Robot Learning; 2017 Nov 13–15; Mountain View, CA, USA; 2017.
[72] Edmonds M, Gao F, Xie X, Liu H, Qi S, Zhu Y, et al. Feeling the force: integrating force and pose for fluent discovery through imitation learning to open medicine bottles. In: Proceedings of 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2017 Sep 24–28; Vancouver, BC, Canada. New York City: IEEE; 2017. p. 3530–7.
[73] Fire A, Zhu SC. Learning perceptual causality from video. ACM Trans Intell Syst Technol 2015;7(2):1–22.
[74] Zhao Y, Holtzen S, Tao G, Zhu SC. Represent and infer human theory of mind for human–robot interaction. In: AAAI Fall Symposia; 2015 Nov 12–14; Arlington, VA, USA; 2015.
[75] Zhu Y, Zhao Y, Zhu SC. Understanding tools: task-oriented object modeling, learning and recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2015 Jun 7–12; Boston, MA, USA; 2015.
[76] Huang SH, Huang I, Pandya R, Dragan AD. Nonverbal robot feedback for human teachers. 2019. arXiv:1911.02320.
[77] Balbach FJ. Measuring teachability using variants of the teaching dimension. Theor Comput Sci 2008;397(1–3):94–113.
[78] Goldman SA, Kearns MJ. On the complexity of teaching. J Comput Syst Sci 1995;50(1):20–31.
[79] Pearl J. Causality. Cambridge: Cambridge University Press; 2009.
[80] Bradley RA, Terry ME. Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika 1952;39(3–4):324–45.
[81] Ramachandran D, Amir E. Bayesian inverse reinforcement learning. In: Proceedings of International Joint Conference on Artificial Intelligence; 2007 Jan 6–12; Hyderabad, India; 2007.
[82] Baker CL, Saxe R, Tenenbaum JB. Action understanding as inverse planning. Cognition 2009;113(3):329–49.
[83] Goodman ND, Frank MC. Pragmatic language interpretation as probabilistic inference. Trends Cogn Sci 2016;20(11):818–29.
[84] Yu X, Han B, Yao J, Niu G, Tsang I, Sugiyama M. How does disagreement help generalization against label corruption? In: Proceedings of the 36th International Conference on Machine Learning; 2019 Jun 10–15; Long Beach, CA, USA; 2019.
[85] Li J, Socher R, Hoi SCH. DivideMix: learning with noisy labels as semi-supervised learning. 2020. arXiv:2002.07394.
[86] Berthelot D, Roelofs R, Sohn K, Carlini N, Kurakin A. AdaMatch: a unified approach to semi-supervised learning and domain adaptation. In: Proceedings of International Conference on Learning Representations; 2022 Apr 25–29; online; 2022.
[87] Yuan L, Fu Z, Shen J, Xu L, Shen J, Zhu SC. Emergence of pragmatics from referential game between theory of mind agents. In: Emergent Communication Workshop, 33rd Conference on Neural Information Processing Systems; 2019 Dec 8–14; Vancouver, BC, Canada; 2019.
[88] Lazaridou A, Peysakhovich A, Baroni M. Multi-agent cooperation and the emergence of (natural) language. In: International Conference on Learning Representations; 2017 Apr 24–26; Toulon, France; 2017.
[89] Lazaridou A, Hermann KM, Tuyls K, Clark S. Emergence of linguistic communication from referential games with symbolic and pixel input. In: International Conference on Learning Representations; 2018 Apr 30–May 3; Vancouver, BC, Canada; 2018.
[90] Watkins CJ, Dayan P. Q-learning. Mach Learn 1992;8(3–4):279–92.
[91] Williams RJ. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Mach Learn 1992;8(3–4):229–56.
[92] Chen X, Cheng Y, Tang B. On the recursive teaching dimension of VC classes. In: Proceedings of the 30th International Conference on Neural Information Processing Systems; 2016 Dec 5–10; Barcelona, Spain; 2016.
[93] Doliwa T, Fan G, Simon HU, Zilles S. Recursive teaching dimension, VC-dimension and sample compression. J Mach Learn Res 2014;15:3107–31.
[94] Mitchell TM. Machine learning. New York City: McGraw-Hill; 1997.
[95] Yuan L, Zhou D, Shen J, Gao J, Chen JL, Gu Q, et al. Iterative teacher-aware learning. In: Proceedings of the 35th International Conference on Neural Information Processing Systems; 2021 Dec 6–14; online; 2021.
[96] Babes M, Marivate V, Subramanian K, Littman ML. Apprenticeship learning about multiple intentions. In: Proceedings of the 28th International Conference on Machine Learning (ICML-11); 2011 Jun 28–Jul 2; Bellevue, WA, USA; 2011.
[97] MacGlashan J, Littman ML. Between imitation and intention learning. In: Proceedings of the 24th International Joint Conference on Artificial Intelligence; 2015 Jul 25–Aug 1; Buenos Aires, Argentina; 2015.
[98] De Weerd H, Verbrugge R, Verheij B. Negotiating with other minds: the role of recursive theory of mind in negotiation with incomplete information. Auton Agent Multi-Ag 2017;31(2):250–87.
[99] Ziebart BD, Maas AL, Bagnell JA, Dey AK. Maximum entropy inverse reinforcement learning. In: Proceedings of the 23rd AAAI Conference on Artificial Intelligence (AAAI); 2008 Jul 13–17; Chicago, IL, USA; 2008.
[100] Vroman MC. Maximum likelihood inverse reinforcement learning [dissertation]. New Jersey: Rutgers University-Graduate School-New Brunswick; 2014.
[101] Liu W, Dai B, Li X, Liu Z, Rehg J, Song L. Towards black-box iterative machine teaching. In: Proceedings of the 35th International Conference on Machine Learning; 2018 Jul 10–15; Stockholm, Sweden; 2018.
[102] Wu L, Tian F, Xia Y, Fan Y, Qin T,Lai J, et al. Learning to teach with dynamic loss functions. In: Proceedings of the 32th Conference on Neural Information Processing Systems; 2018 Dec 3–8; Montreal, QC, Canada; 2018.
[103] Gao X, Gong R, Zhao Y, Wang S, Shu T, Zhu SC. Joint mind modeling for explanation generation in complex human–robot collaborative tasks. In: Proceedings of 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN); 2020 Aug 31–Sep 4; Naples, Italy; 2020.
[104] Yuan T, Liu H, Fan L, Zheng Z, Gao T, Zhu Y, et al. Joint inference of states, robot knowledge, and human (false-) beliefs. In: Proceedings of 2020 IEEE International Conference on Robotics and Automation (ICRA); 2020 May 31–Aug 31; Paris, France; 2020.
[105] Yuan L, Gao X, Zheng Z, Edmonds M, Wu YN, Rossano F, et al. *In situ* bidirectional human–robot value alignment. Sci Robot 2022;7(68):eabm4183.
[106] Russell S. Human compatible: artificial intelligence and the problem of control. New York City: Viking; 2019.
[107] Tang N, Stacy S, Zhao M, Marquez G, Gao T. Bootstrapping an Imagined We for cooperation. In: Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci); 2020 Jul 29–Aug 1; online; 2020.
[108] Stacy S, Zhao Q, Zhao M, Kleiman-Weiner M, Gao T. Intuitive signaling through an "Imagined We". In: Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci); 2020 Jul 29–Aug 1; online; 2020.
[109] Bara CP, Ch-Wang S, Chai J. MindCraft: theory of mind modeling for situated dialogue in collaborative tasks. In: Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP); 2018 Nov 2–4; Brussels, Belgium; 2018.
[110] Fan L, Qiu S, Zheng Z, Gao T, Zhu SC, Zhu Y. Learning triadic belief dynamics in nonverbal communication from videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA; 2021.
[111] Arora S, Doshi P. A survey of inverse reinforcement learning: challenges, methods and progress. Artif Intell 2021;297:103500.
[112] Blumer A, Ehrenfeucht A, Haussler D, Warmuth M. Learnability and the Vapnik–Chervonenkis dimension. J ACM 1989;36(4):929–65.
[113] Bartlett PL, Bousquet O, Mendelson S. Localized Rademacher complexities. In: Proceedings of International Conference on Computational Learning Theory; 2022 Jul 2–5; London, UK; 2002.
[114] Chapelle O, Schölkopf B, Zien A. An augmented PAC model for semi-supervised learning. In: Chapelle O, Schölkopf B, Zien A, editors. Semi-supervised learning. Cambridge: MIT Press; 2006.
[115] Barbu A, Pavlovskaia M, Zhu SC. Rates for inductive learning of compositional models. In: Workshops at the Twenty-Seventh AAAI Conference on Artificial Intelligence; 2013 Jul 14–18; Bellevue, WA, USA; 2013.
[116] Hintikka J. Knowledge and belief: an introduction to the logic of the two notions. Stud Log 1962;16:119–22.
[117] Aumann RJ. Agreeing to disagree. Ann Stat 1976;4(6):1236–9.
[118] Halpern JY, Moses Y. Knowledge and common knowledge in a distributed environment. J ACM 1990;37(3):549–87.
[119] Smith NJ, Goodman ND, Frank MC. Learning and using language via recursive pragmatic reasoning about other agents. In: Proceedings of the 26th

International Conference on Neural Information Processing Systems; 2013 Dec 5–10; Lake Tahoe, NV, USA; 2013.
[120] Carston R. Informativeness, relevance and scalar implicature. Pragmat Beyond New Ser 1998;37:179–238.
[121] Vogel A, Bodoia M, Potts C, Jurafsky D. Emergence of Gricean maxims from multi-agent decision theory. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2013 Jun 9–14; Atlanta, GA, USA; 2013.
[122] Turing AM. On computable numbers, with an application to the Entscheidungsproblem. Proc Lond Math Soc 1937;2(1):230–65.
[123] Stone P, Kraus S. To teach or not to teach? Decision making under uncertainty in *ad hoc* teams. In: Proceedings of the Ninth International Conference on Autonomous Agents and Multiagent Systems; 2010 May 10–14; Toronto, ON, Canada; 2010.
[124] Zhang A, Sodhani S, Khetarpal K, Pineau J. Learning robust state abstractions for hidden-parameter block MDPs. In: Proceedings of the International Conference on Learning Representations; 2020 Apr 26–May 1; online; 2020.
[125] Barrett S, Rosenfeld A, Kraus S, Stone P. Making friends on the fly: cooperating with new teammates. Artif Intell 2017;242:132–71.