Contents lists available at ScienceDirect

Engineering

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

Research
Artificial Intelligence—Review

通讯式学习——统一的机器学习模式

袁路遥[a,b,*]，朱松纯[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

关键词
人工智能
合作沟通
机器学习
教育学
心智理论

摘要

在本文中，我们提出了通讯式学习的框架，从而统一已经存在的机器学习范式，如被动学习（passive learning）、主动学习（active learning）、算法式教学（algorithmic teaching）等；同时促进新的学习方法的发展。扎根于人类的合作式通讯，这个范式用通讯的过程来刻画学习，同时将传授（pedagogy）的思想应用在机器学习领域。引入传授让机器可以更好地利用多种信息源进行学习：除了传统的随机抽样数据，还包括来自于因材施教的老师有目的性的信息。具体来讲，在通讯式学习模式中，一个老师和一个学生通过交流完成特定知识的学习过程。每个智能体都有一套思维（mind），包括其知识（knowledge）、效用（utility）和思维的变迁规则（dynamics）。每个智能体同时估计其伙伴的思维以进行高效的交流。我们给出了可以支持这种递归过程的师生思维表征（mental representation）和学习过程的公式（learning formulation），这些结构让通讯式学习具有和人类相似的学习效率。我们进一步用一些典型的人机合作任务来展示通讯式学习模式的可行性，同时说明了这个模型可以超越香农（Shannon）的通讯极限。最后，我们给出了通讯式学习框架对于基础学习理论的贡献，包括提出了学习的阶层以及定义了学习的停机问题。

1. 引言

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。——中国谚语

When I walk along with two others, they may serve me as my teachers.

三人行，必有我师焉。——孔子

1.1. 以统一的学习模式为目标

统计和机器学习近来的快速发展让人工智能（AI）能够在诸多特定的任务上取得出色的表现。然而，当前的机器学习范式同时也显现了一些不足：需要大量的训练数据，不可解释或共享的表征，对于新任务和未知场景的泛化性不足。这些机器学习方法属于“大数据，小任务”的范式[1]，与能够用很少的数据完成诸多日常生活任务交流的（小数据，大任务）人类学习方式有着天壤之别。人类学习同时还牵涉多层的认知机理，并且是构建在包括多

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

英文原文：Engineering 2023, 25(6): 77–100
引用本文：Luyao Yuan, Song-Chun Zhu. Communicative Learning: A Unified Learning Formalism. *Engineering*, https://doi.org/10.1016/j.eng.2022.10.017

个参与者沟通的基础上，典型的情景就是老师向学生传授的过程。可惜的是，人类学习的此种精妙性通常被现在的机器学习算法简化了。为了填补机器学习和人类学习存在的差距，在本文中我们提出“通讯式学习”（以下简称“通学”）模式。通学可以整合多种学习范式，它把用一个老师和一个学生的交流过程来刻画学习。在通学模式中，每个智能体的思维表征包含以下结构。

• 自我思维：包括这个智能体目前对于知识的信念（belief of the knowledge）、效用（utility）以及思维变潜规则（dynamic functions）。

• 镜像思维：对于伙伴自我思维的估计。

• 共识（common mind）：已经在伙伴间形成了的共同知识。

• 上帝思维（God’s mind）：关于客观世界的事实。

这些思维的组成部分会共同驱动学习和交流的过程。立足于通学模式，我们用通讯过程来刻画学习并展示这种模式相对于传统的学习模式的优势。进一步地，我们发现通学能够用一种模式涵盖并超越已有的机器学习方法。我们从老师学生协作通讯的视角出发，综述了多种目前常见的机器学习算法。我们发现这些算法都可以表示为通学的特例。所谓千举万变，其道一也，通学为读者们提供了一个理解多种多样的机器学习算法的统一模式，并且该模式能够与协作性的教学（cooperative pedagogy [2]）有机结合来产生更多更高效的学习算法。

1.2. 通讯式学习的认知机理

在人类社会中，通讯交流的行为随处可见，以至于我们经常对其先进、精妙习以为常，忽视了支持这一类行为的认知机理（cognitive infrastructure）的复杂性。事实上，即便原始如有目的地发信号（intentional signal）这种简单的交流行为，在自然界中也极其罕见，已知只有灵长类（primates）甚至人科动物（great apes/Hominidae）可以完成[3]。完整的人类通讯交流系统是通过构建共享注意力（joint attention）和共识（common ground）来达到协作目的的复杂系统。人与人之间的合作规范（cooperative norm）、沟通交流传统（communication convention）和能够进行递归推理（recursive reasoning）的认知机理（cognitive infrastructure）共同使得这个系统可以顺利且高效地发挥作用[3]。学习作为一个终身都在进行的交流过程也通过这个系统完成。其精妙、高效和复杂都是人类智能的前提条件，也是当今人工智能想要复现的功能。

最近十年来的认知心理学[4–5]和人类学[3]研究揭示了人类的交流和学习都是建立在多层次的认知结构和通讯协议上的。为了解释这种复杂性，我们给出在通学模式中的智能体A和B的思维表征如图1所示。这两个智能体都可以是人工智能或者人类，并且老师和学生的角色是对称的，师生角色是可以随着交流的进行而切换的。这套表征有如下的特性。

（1）心智理论（theory of mind[†] [6]）。每个智能体具有推己及人地估计伙伴当前的思维的能力：

• G：上帝的思维。这个思维包含了世界的真实状态并且会根据一定的模型（world transition model）演化。

• P_t：老师A的思维。包含了其关于世界的知识、效用、行为方式和思维变迁模型。

• Q_t：学生B的思维。包含了其关于世界的知识、效用、行为方式和思维变迁模型。

• $\widehat{Q}_t$：老师认为学生知道什么。这个思维帮助老师进行协同性的教学（cooperative pedagogy）。

• $\widehat{P}_t$：学生认为老师知道什么。这个思维帮助学生更好地理解老师的信息。

• C_t：师生间的共识。每个人都知道；每个人知道对方知道；每个人知道对方知道自己知道，以此类推[7]。

这6个思维间的区别和距离让老师和学生间进行有目的的信息交换而非随机的数据抽样，从而驱动着通讯和学习的过程。

（2）适应性的共识和学习协议（learning protocols）。共识C_t是未来通讯和学习的基础。师生间的共识越多，他们学习的效率越高。例如，他们可以通过类比来完成新的概念学习（concept learning [4]），并且通过微调已存在于共识中的概念来获得新的概念。设想如果学生已经学过“狗”这个概念，那么当老师要传授“狼”这个概念时，她就可以只类比狗和狼的区别而不必重复二者的相同性。通学的思维表征让共识得以积累并且应用于产生更高效的学习协议。

我们发现，当具备了满足上述条件的合适认知机理后，通学模式可以为机器学习领域带来如下的贡献。

• 提出了统一的学习框架，在多个维度涵盖了已有的机器学习方法为其特例：监督和非监督、主动和被动、因果实验和因果观察（observational causality）等。并且简化了在未知情境下发展出新的学习协议的过程。

• 将教学法（pedagogy）和机器学习相结合，在协作通讯的情况下，通过考虑因材施教的老师来加快学习。

• 拓展了已有的统计学习和通讯极限，提出可以超越

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

如香农极限（Shannon's communication limit [8]）以及瓦利安特的 Probably Approximately Correct (PAC) Learning [9]的统计学习极限的学习协议。

• 研究了机器学习的基本问题和定义了学习的停机问题。

一言以蔽之，通学模式使得老师可以根据多种不同的标准选择对于学生最有帮助的信息[10–12]，同时学生根据他对于老师传递信息的机制的估计来更快地理解知识[13–15]。根据老师不同的信息选择机制，我们可以将多种机器学习算法统一在通学模式下，从简单的被动学习到相对复杂的协同教学。后者给出的学习协议能够达到显著优于传统机器学习理论所刻画的极限，因为对于学生而言数据不再是来自于所及抽样，而是有目的的老师。在第4节我们会详述通学模式以及其数学定义。

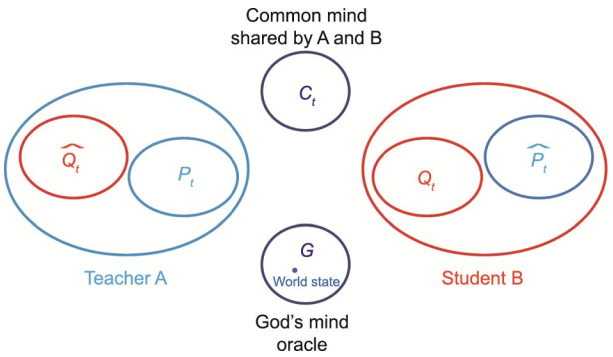

图1. 同学的关键表征包括6个随时间迁移的思维。G：上帝的思维表示客观世界状态；P_t：老师A的思维；Q_t：学生B的思维；$\widehat{P}_t$：学生对老师的估计；$\widehat{Q}_t$：老师对学生的估计；C_t：师生间的共识，A和B都知道对方知道。

1.3. 机器学习的统一框架

通学模式可以作为一个统一的框架来整合现有的机器学习算法，将他们表示为自己的特例。在20世纪60年代，机器学习被首次提出用于让计算机可以进行模式识别（pattern recognition [16–18]）。时至今日，随着数据和计算资源的丰富，各种机器学习算法在诸多领域都很好地完成了问世之初理解数据的目的。复杂的模型被应用于从图像分类[19–20]、物体检测[21–23]、句子生成[24]到超越人类的游戏竞技[25–26]等各种任务中。

然而大部分流行的机器学习算法着重于优化个体的学习者（individual learner），完全依赖于来自单智能体的经验。这些经验通常来自于马尔科夫决策过程（Markov decision process [25–26]）、从数据分布中的随机抽样[19, 27]、专家（oracle）对于查询（query）的回复[28–29]，或者来自于专家（expert）的演示（demonstration）[30]。直到最近，研究人员才意识到，在贝叶斯概念学习（Bayesian concept learning [13,31–35]）和从示范中学习（learning from demonstration [15,36–37]）领域，引入进行传授的老师（pedagogical teacher）可以实现优于从随机数据或者从来自于专家的最优演示中学习的效果。同时，机器教学（machine teaching [38–42]）算法也开始在连续参数空间（continuous parameter space）和大数据问题中对合作性的老师建模。在图2中我们对比了典型的学习范式。

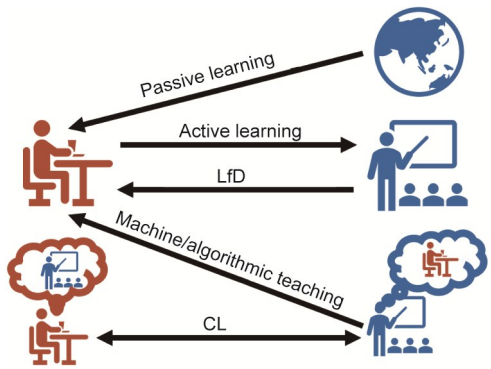

图2. 不同学习和教学协议的对比。蓝色表示老师，红色表示学生。气球表示对于伙伴的估计。

1.4. 将教学法融入机器学习

通学的另一个优势是将人类学习的教学法融入了机器学算法的开发之中。生活中，当我们提到学习时，不管想到的是咿呀学语的孩童亦或是学校里的学生，场景里总是至少有两个人——一个老师和一个学生。老师试图将她的知识通过最有效的手段，利用最有帮助的材料传授给学生；学生则主动地吸收来自老师的信息以尽快达到学习目标。这种发生在多智能体系统中的学习就是我们所说的教学法（pedagogy [2]）。然而，在一般的机器学习中，老师的角色通常会被随机生成的训练数据[以数据集（dataset）或者交互经历（interactive experience）的形式出现]代替[18]。机器学习算法就好比我们人类社会中的科学家，试图通过观察自然界中随即发生的现象，比如苹果下落、点点星光、刮风下雨，来找到规律，解释自然现象。然而最常见的学习并不是以科学发现呈现出来的。人类社会中的学习大部分是通过师生的沟通交流、协作性的口传心授完成的信息传递和知识累积[42,45]。

在认知科学领域有丰富的实验证据表明人类在传授和学习的情境下有能力而且有倾向进行更高效的学习。从婴儿（infants）和幼童（toddlers）可以区分抽样的随机与否[46]到他们对于传授行为（pedagogical behavior）的意识[47–49]再到儿童和成人会运用教学法推理（pedagogical inference）来促进单词的学习（word learning）[50]，这些

研究结果都说明人类在学习过程中对于随机产生的信息和教学信息的反应很敏感，而且可以更高效地利用后者促进学习。人类从很小的时候起就可以从教学传授的场景[31]中获得额外的信息，并且他们还知道，如果自己是老师的话该如何利用这些额外信息来帮助别的学生[51]。

科学家和学生的学习方式的本质区别在于后者有一个因材施教的老师，经过她挑选之后发给学生的信息比随机产生的信息对于学习更有帮助，让通讯的效率大幅提升。具体来说，老师会根据不同的学生因材施教；同时学生在熟悉了老师的教学方式之后可以通过推测老师在发送信息时的动机，从而学得更快[10,18,48,52–53]。最近，研究者们逐渐意识到传统机器学习算法对于数据的过大需求和“大数据，小任务”[1]的局限，特别是和人类利用有限数据就可以高效学习相比的不足[31,48,54]。因此，通过结合教育学（pedagogy）以优化机器学习的研究方兴未艾[13,15,31–42]。然而，即便如此，相比于人类学习，这些工作对于学生的建模还不够复杂准确，不足以完全理解老师的教学法的全部意图，甚至与被动地从数据中学习别无二致。在文献[45,55]中研究者们在师生模型中加入了递归推理能力，从而实现了心智理论[6]。然而，其中的分析实验主要关注在贝叶斯概念学习的情境下，只牵涉比较有限的数据和假设空间（hypothesis space）。在3.3节中，我们将介绍通学模型如何将教育学和机器学习相结合以开发更加先进的学习算法。本文以下内容的结构为：在第2节中，我们会阐述学习和通讯的关系并且指出定义通学模式的深层动机。在第3节中，我们将系统地介绍通学模式。从每个智能体的建模，其思维表征到具体学习迁移函数的数学定义。我们还将在3.4节中说明现有的机器学习算法都可以表示为通学模式的特例（更多细节见附录A中的S1节和S3节）。接着在第4节中，我们会用一些具体实例展示通学模式的实用性和必要性。我们将带来一个指代游戏和一个人机合作的任务作为实例研究。随后第5节给出了通学模式对于机器学习基础理论的贡献。在5.1节中，我们会提出一个对于学习过程的新表示并且给出一个超越香农极限的学习协议。在5.2节中，我们会厘清机器学习的三个阶层并且定义学习的停机问题。最后本文将以通学模式的影响结束。

2. 以通讯的视角审视学习

让我们回到科学家和学生两种学习方式的对比。对于大部分曾经是或者依然是学生的人来说，毋庸赘言学习是发生在交流的过程中。学生通过和老师互相传递信息来完成知识、技能和价值观的学习。在西方文化中，这种形式被认为起源于苏格拉底。传说他经常用对话的方式传播自己的思想。事实上，从大自然到科学家的信息传递也可以被视为一种通讯，只要我们承认有一个全知的存在把思维中的知识以自然现象为信息传递给科学家。知识这种信息微言大义，对大数人来说远比苏格拉底的信息晦涩难解。更普遍的来说，我们可以将学习解读为将信息从一个思维中传递到另一个思维的过程。有趣的是，这正和通讯的标准定义相吻合。那么通讯和学习这两个紧密联系的概念具体的关系是什么呢？要回答这个问题，我们必须要从信息理论（information theory）[8]和统计学习（statistical learning）理论[56]谈起。这两套理论体系分别从数学上严谨地描述了通讯和学习过程。我们将发现二者具有紧密的联系。

2.1. 通讯和学习的联系

1948年克劳德·香农提出信息理论[8]作为研究通讯（如电话信号传输）的标准框架。如图3所示，在这个框架中传输者（sender）和接收者（receiver）有一个共享的密码本（shared codebook）。信息是用于描述外部世界的特定状态（world state），如关于室内场景（indoor scene）的解析图（parse graph），或者一段话的语义（semantics）。这个通讯系统的目的是为了最大限度地在接收者初处还原发送者的信息。简而言之，它包括了一个信息源从信息空间中选择一个信息，并将其编码（encode）为适合传输的一系列通讯符号（communication symbol）。这一系列符号接着会通过有噪声的介质（noisy medium）传递给接收者解码（decode）并重建原始的信息。香农的主要贡献是推导出了信息传递速率[如比特每秒（bit per second）]的上限和传输介质之间的关系。这个速率的极限就被称为信道容量（channel capacity）。也就是说，每次信息传输到接收者之后，接收者对于信息所描述的世界状态的不确定性就会减少，减少的幅度由信道容量决定。例如，我们用w^t表示t时刻所有可能的世界的集合。那么通过t+1时刻的信息所能获得的信息增益（information gain）可以表示为：

$$\mathrm{IG}^{t+1}=\log_2\frac{1}{|\mathcal{W}^{t+1}|}-\log_2\frac{1}{|\mathcal{W}^{t}|}=\log_2\frac{|\mathcal{W}^{t}|}{|\mathcal{W}^{t+1}|} \quad (1)$$

香农理论的一个不足在于它缺乏对于信息的语义或含义的描述。发送者和接收者虽然假设彼此有一致的共识，如共享的密码本，但是通讯协议不考虑收发者的思维状态或者协作动机。以此计算出的通讯效率上限必定和团队合作中的沟通速率有不符之处。

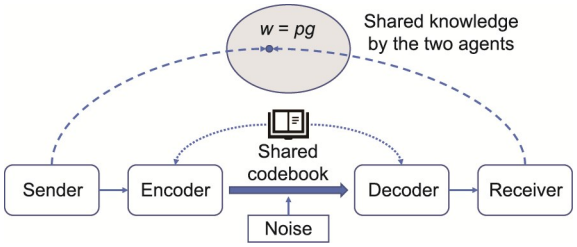

图3. 香农通讯理论的示意图。共享的密码本起到了协作通讯中共识和通讯规范的部分作用。

2.2. 统计机器学习

相比于信息理论，问世更晚的统计学习理论却更进一步。如在众所周知的概率近似正确学习[probably approximately correct (PAC)]模型中，莱斯利·瓦利安特把学习建模为准确地根据训练数据还原出未知的假设[9]。图4展示了PAC-learning的情境。严格来说，学习者试图通过随机抽样与外部世界的例子学习一个概念c，这个概念可以是在状态空间的一个集合Ω [9]或者是假设空间Θ ($\theta \in \Theta$)中的一个概率模型θ。随机抽样的例子可以表示为$\{(I_i, c_i), i=1, \cdots, M\}$。学习是由一个先定的（predefined）效用（utility）或损失（loss）函数u驱动的。PAC-learning理论给出了需要学到误差$\leqslant\epsilon$、确定性$\geqslant\delta$的训练数据的个数下限$n(\epsilon, \delta)$。

因此，只要将目标概念等同于信息，将训练数据等同于通讯符号我们就可以发现香浓通讯模型和PAC-learning统计学习模型的关联性。我们可以回到之前的比方，具体来说：

- 大自然像信息源一样从概念空间中按照一定的概率选择一个概念作为信息，接着用一系列的训练数据作为通讯符号来为这个概念进行编码；
- 编码过程中这些训练数据的标注（label）可能会由于噪声等原因被随机污染改变（corrupted）；
- 在收到有噪声的训练数据之后，学习者试图解码并还原出原始的概念[58]；
- 就像信道容量之于香农通讯，在统计学习中，想要实现足够确定和低误差的概念传输所需要的训练数据量也有一个下限，即样本复杂度（sample complexity）。

总的来说，统计学习理论和信息理论的通讯框架是相似的，只是更加强调解码的过程。顺理成章地，统计学习理论也有和信息理论一样的不足，即没有对于智能体协作性和思维状态的建模。如果我们可以克服这一不足，我们是否可以突破PAC-learning给出的样本复杂度极限呢？本文提出的通学模式就试图回答了这一问题。在这个模式中，信息是由老师根据心智理论有目的地选择而来，携带不止于字面意思的额外含义来辅助学生的学习。发展出来的通讯协议也更高效，允许师生互相听懂对方的“弦外之音”。

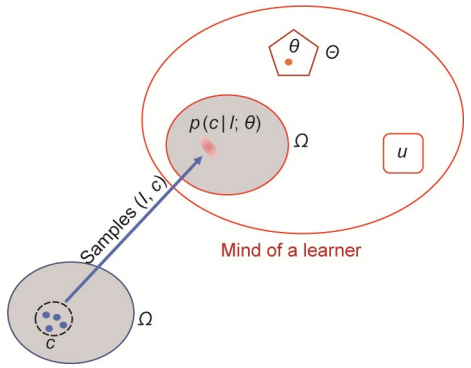

图4. 统计学习理论的示意图。学习者被动地接受来自于物理环境的例子。学习者的思维不包含心智理论需要的结构。

2.3. 超越香农和Valiant的框架

在我们回顾了信息理论和统计学习理论，特别是PAC-learning的基础和联系之后，敏锐的读者一定已经意识到瓦利安特的学习模型除了应用在“科学家”的学习场景以外，对于师生合作学习的场景只有很有限的解释力。因为这个学习模型假定学生收到的数据都是随机抽样得到的，然而在师生合作中，学生收到的信息是来自于老师在考虑学生当前思维状态之后有目的性地发出的。同时，由于老师角色的缺失，学生在学习中只需要被动地接受数据，而不会主动推测信息的来源是否承载着额外的对于学习有益的信息。这和我们之前提到的人类学习方式大相径庭。在人类学习中，我们不仅会理解信息的字面意思，还会主动根据疑惑提问，并且（通过心智理论）基于对于伙伴的了解分析他们发出特定信息而非其他信息的动机，预判他们收到某些信息的反应并从中解读出额外的含义。

这里，我们用一个简单的例子来说明有目的性的信息相对于随机数据所具有的优势。假设有四个数字集合{1,2,3},{4,5,6},{4,5},{1,2,4,6}，老师可以通过发送属于目标集（target set）的数字来告知学生这四个集合中的哪一个是目标集。我们可以考虑第三组{4,5}为目标的情况。对于一个从随机抽样中学习的学生来说，需要许多重复的数字来塑造他的信念并排除第二组，而第二组恰好是目标的超集。学生只有在不断地没有接收到第二组中的数字6之后，才能够概率性地排除第二组数字，但仍然不能确信目标集是{4,5}。

然而，如果向合作的教师学习，智能体之间可以建立更高效的协议，有些甚至可以只用一个信息就完成目标的确定识别。例如，一个合作的教师将使用3来唯一地指代

第一组，因为没有其他组包括3，这样1和2就可以用来指代第四组，因为如果目标是第一组，教师会选择3。同理，从第四组中将6解放出来，可以用其表示第二个组。剩下4和5两个都可以没有歧义地指代第三组（见图5[60]从随机数据中进行的统计学习与这种教学学习之间的比较）。这样的学习协议只有在合作性和教学性的智能体之间才有可能。

因此通过对话进行的苏格拉底式的学习模式是超越PAC-learning，即单向的概念通讯。它涉及意义和意图。其实这并不是一个完全新颖的观点。正如香农本人所承认的那样[8]，“信息经常是有意义的；也就是说，根据某些系统，它们指的是或与某些物理或概念实体相关联。通信的这些语义方面对于工程问题是不相关的。”在后续的工作中，沃伦·韦弗（Warren Weaver）也指出对通讯的更广泛的理解应该包括通讯参与者间的思维影响过程[60]。具体来说，他建议从三个层面考虑普遍的通讯问题：

（1）通讯的符号如何准确地传递（技术问题）；

（2）被传达的符号如何精确地传达所需的意义（语义问题）；

（3）传达到的意义如何有效地影响行为（有效性问题）。

也就是说符号传递只是为通讯过程中的参与者提供必要技术基础，以保证后续层面的通讯。事实上，人类的交流要复杂得多。为了说明语义通讯（第二层次）是如何与传递的符号（第一层次）相独立的，我们不妨想象一下人们对同一电视图像的不同反应。例如，敌对双方的球迷对一场足球比赛的理解。此外，即使对某一特定信息只有一种理解，在人类交流中，接收者也通常不会简单地接受，而是可以选择忽略或反对某一信息（第三层次）。美国关于气候变化的激烈的政治讨论就是这种现象的一个典型例子。

由于我们的目标是用通讯模型来解释学习，根据之前的工作，我们需要着眼于通讯的两个更高层次，主要是语义层次进行说明。为了理解第二层次的通讯是如何在人类之间进行的，我们参考了迈克尔·托马塞洛（Michael Tomasello）从进化论和认知的角度所做的说明[3]。他认为，人类交流成功的关键是参与者的共识（common ground），其中包括共同的注意力、共同的经验和共同的文化知识。共识提供了关键的情境、社会、政治、文化和历史背景，使人们能够根据所收到的通讯符号构建意义。

我们用文献[3]中的一个例子来说明这一点。假设我们正在去图书馆的路上，我突然指向靠在图书馆墙上的自行车。你的反应多半是很疑惑的，因为你不知道我想表达什么意思，因为以手指物本身而言，没有任何意义。但是，如果几天前你刚刚和你的伴侣很不愉快的分手，而且我们都知道这件事，并且那辆自行车是他的，我们也都知道这一点，那么刚才的指向性手势就可以传达非常复杂的内容，比如“你的男朋友已经在图书馆了（所以也许我们是不是不进去了）”。或者，如果那自行车是我们共同知道你刚刚被偷的，那么完全相同的指向性手势将意味着截然不同的东西。再或者，我们一直担心这么晚了，图书馆是不是还开着，然后我指向外面的自行车，以此说明图书馆还开着，等等情况不一而足。

需要指出的是，共识并不是一个新的概念，因为它已经存在于香农的通信模型和Valiant的PAC-learning中了。在传统的通信中，它由发送方和接收方共享的密码本表示，而在PAC学习中，它是共同的概念空间。只是在人类通讯不一样，共识是由参与者共同构建的，而不是由建立通讯系统或学习算法的第三人赋予的。

以上发现表明了学习和通讯之间的联系，并指出了通学模式的必要性。在下一节中，我们将展示老师和学生的思维表征，以及他们如何通讯以实现高效的教学法。

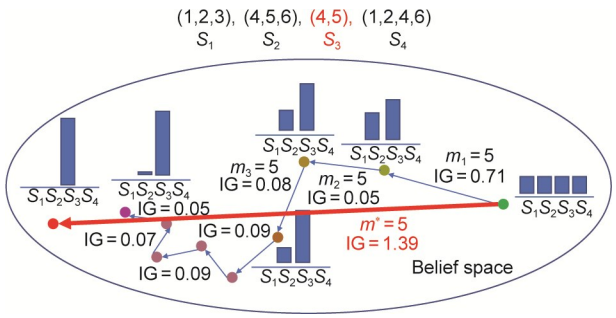

图5. 学生从随机抽样（蓝色轨迹）和合作老师（红色轨迹）中学习时的信念更新过程。{4,5}具有单个消息，因为该概念空间的递归教学维度（recursive teaching dimension, RTD）[59]是数字1，它是常规教学维度（teaching dimension, TD）的下界。具体地说，学生知道老师会发送数字3代表S_1；因此，数字1或2表示S_4。类似地，数字6只能表示S_2，使5成为S_3唯一标识符。这种递归协议的先决条件是老师是教学性的，并且学生意识到她的帮助。m_1、m_2和m_3表示第一、第二和第三条随机消息，m^*表示来自合作老师的消息。更多具有不同TD的概念类见附录A中的表S1。

3. 通讯式学习

在本节中，我们将给出通学模式的定义。我们先从它的基础结构和老师、同学的思维表征表示开始介绍。正如我们在1.2节所讨论的，通学模式的基础结构需要有足够的表达能力才可以支持共识和心智理论的建模。接着，我们在3.3节中介绍学习的动态函数（dynamics），说明在整合了合作推理和教学法之后学习是如何进行的。

3.1. 完整的师生思维表征

在前几节中，我们指出了心智理论是通学模式的前提条件，但没有严格地定义什么是思维。现在我们详细阐述一下通学模式中的思维表征。在通学模式中，有两个智能体——老师和学生，他们之间有六套思维。第一对是智能体的自我思维（egocentric minds）。第二对是智能体对其伙伴的思维的理解和估计。再加上共识和上帝的思维†，也就是客观事实的表述，一共构成六套思维。理论上说，多智能体系统（multiagent system）中的相互推理是递归的，如果不存在收敛，可以无限进行下去[7,61–64]。例如，A知道事实e；B知道A知道事实e；A知道B知道A知道事实e，如此循环往复。为了避免计算上的困难，我们只建立了一级递归的模型，这也和多项人类研究中发现的人类认知能力相匹配[65–66]。另外要注意的是，本节中的表述不包括学习的动态函数，即在智能体之间的互动过程中，这些思维如何更新。在3.3节中，我们会详细地讨论这部分内容。现在，我们先假定老师和学生有办法根据来自世界和伙伴的信息更新自己的思维。我们说一个智能体的思想包括以下组成部分。

（1）当前情况的状态表示 $\boldsymbol{w}$。w代表当下的世界，通常是一种有内在结构的思维状态。例如，一个状态w=*pg*是一个解析图，在更一般的情况下，w是一个时空因果[space, temporal & causal (STC)]解析图，STC-pg这种表征已经被我们广泛应用在计算机视觉[67–68]、语言理解[69–70]、机器人学[71]和常识推理[72–75]中。在通学模式中，信息将以情境通讯的模式交换，也就是解析图代表了一个场景的构成，如一个客厅和其中的物体，以及动作及其导致的流变（fluents，物体状态随时间的变化）。如图6所示，一个解析图由三个部分组成：

- $pg_{[0,\ t_0]}$总结了目前的情况（图6中的蓝色）；
- $pg_{[t_0,\ t_0+t_\Delta]}$预测意图和计划（图6中的绿色）；
- $pg_{[t_0,\ t_0+t_\delta]}$是当前的注意力（图6中的红色三角形内）。

由于通学是一个迭代的过程，解析图将随着时间的推移在多个语义层面通过消息进行交流。通学智能体有一个共识，其状态w_c=pg_c（图7）也包括$pg_{c[0,t]}$、$pg_{c[t,t+t_\Delta]}$以及$pg_{c[t,\ t+t_\delta]}$，分别表示共享的情境、目标/意图和注意力。正如人类学和认知研究[3]所指出的，这种表征是人类通讯的关键。它使得智能体讯速地在合理的程度理解细节。因此，通学是一个情境学习和通讯的过程，它可以满足复杂的合作通讯需求，比现有的学习方法更加一般化。一些通学模式的简单应用已经在人机交互和团队合作中证明了其可行性和必要性[70–71,75]。

- 空间层次结构：场景-对象-部分-原始结构，用于解析场景和对象；
- 时间层次结构和组成：事件-行动-用于分析事件的动作；
- 因果层次：因用于因果推理和任务计划的行动和线索。

STC−AOG可以被看作是计算机视觉中的图像解析、自然语言理解（natural language understanding, NLU）中的语言解析、机器人技术中的任务规划和常识性人工智能中的认知推理的统一表示。

（2）模型$\boldsymbol{\theta\in\Theta}$属于一个假说/模型空间$\boldsymbol{\Theta}$中。在确定的情境中，如PAC-learning，一个模型是由一个集合代表的概念，如一个物体类别。在有随机性的情境中，模型定义了状态空间上的概率分布$p(s, \theta)$，通常指的是分布的参数。它可以是支持向量机（support vector machine, SVM）的超平面、深度神经网络（deep neural network, DNN）的权重或随机语法的规则。当状态w=pg是一个具有不同构型的结构化解析图时，我们用与或图（and-or graph, AOG）来表示模型$p(w; \theta)$ [67]，这个模型的语言是所有有效构型与其概率的集合。一个解析图pg是语法或AOG的一个实例和实现。如果完整定义的话，一个模型由一个整合了三个层次的STC-AOG表示。

- 空间层次结构：场景—物体—部件（parts）—视觉基元（primitives），用于解析场景和物体。
- 时间层次和构成：事件—动作—动作的分解行动（movements）。
- 因果层次：因果推理和任务规划所需的动作和流变。

STC-AOG可以被看作是计算机视觉中图像解析、自然语言理解中的语言解析、机器人技术中的任务规划和常识性人工智能中的逻辑推理的统一表征。

（3）信念和信念的信念（belief over belief）。我们用I_{A}表示老师的观察和输入，用I_{B}表示学生的观测，I_{A}和I_{B}可以是一个输入图像或视频。我们还用d^t表示在时间t上来自老师的信息，用m^t表示t时刻来自学生的信息。另外，让$a_{\mathrm{A/B}}$表示智能体A/B的行动。由于物理世界和其他智能体的思想对另一个智能体来说都不是完全可以观察到的，所以大多数时候，智能体需要用信念来刻画对于世界和伙伴的理解。将学生到时间t为止的历史表示为：

$$h_{\mathrm{B}}^t=\left[I_{\mathrm{B}}^{1:t}, d^{1:t}, m^{1:t}, a_{\mathrm{B}}^{1:t}\right] \tag{2}$$

† 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.

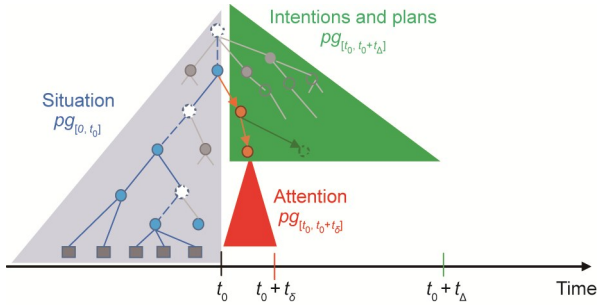

图6. 状态 $w=pg$ 随着时间的推移而展开。t_0 表示当前时间，t_δ 和 t_Δ 分别表示短时间和长时间。$pg_{[0,\,t_0]}$：现状；$pg_{[t_0,\,t_0+t_\delta]}$：当前关注事项；$pg_{[t_0,\,t_0+t_\Delta]}$：意图和计划。

其中，$d^{1:t}$ 表示 $\{d^1, d^2, \cdots, d^t\}$；$m^{1:t}$ 表示 $\{m^1, m^2, \cdots, m^t\}$。

然后，对状态和模型的信念是 $b^t_{\mathrm{B},w}(w)=p(w|h^t_{\mathrm{B}})$ 和 $b^t_{\mathrm{B},\theta}(\theta)=p(\theta|h^t_{\mathrm{B}})$。对老师也是如此，只是通常我们假设 $b^t_{\mathrm{A},\theta}(\theta)=b^0_{\mathrm{A},\theta}(\theta)$；也就是说，老师一开始就有一个正确的模型，且不随时间改变。由于心智理论，师生间也会互相估计对方的信念，这个过程中不确定性也是存在的。因此，我们定义信念的信念为：

$$\mathrm{bob}^t_{\mathrm{AinB},\theta}\left(b_{\mathrm{A},\theta}\right)=p(b_{\mathrm{A},\theta}|h^t_{\mathrm{B}}) \tag{3}$$

$$\mathrm{bob}^t_{\mathrm{AinB},w}\left(b_{\mathrm{A},w}\right)=p(b_{\mathrm{A},w}|h^t_{\mathrm{B}}) \tag{4}$$

其中，$b^t_{\mathrm{A},w}$ 和 $b^t_{\mathrm{A},\theta}$ 是 t 时刻状态和模型 A 的信念；$\mathrm{bob}^t_{\mathrm{AinB},w}$ 和 $\mathrm{bob}^t_{\mathrm{AinB},\theta}$ 是对 t 时刻老师信念 B 的估计。对于老师也是同理。

有的读者可能注意到，随着时间的推移，历史会呈指数级增长。因此在一般情况下，信念的信念是难以精确求解的。在实际操作中，一般会引入一些独立性假设或归纳偏置（inductive bias）来简化计算。对于嵌套信念，因为用贝叶斯滤波器所计算的信念更新是确定性的（deterministic），如果状态空间不是太大，一些近似方法，如粒子滤波器（particle filter）可以处理信念的信念[62]。但是当状态空间很大，甚至是不可数（uncountable）的时候，它的局限性就显现出来了，这是 bob_θ 的一个常见情况。在这些情况下，我们可以通过采取最大后验（maximum a posteriori, MAP）作为分布的近似值来对信念的信念建模[53,76]。

（4）策略函数 π: $W\mapsto\Delta(\mathcal{A})^\dagger$ 从当前状态映射到动作的分布 $a\in\mathcal{A}$。在更普遍的情况下，动作可以组成结构化的计

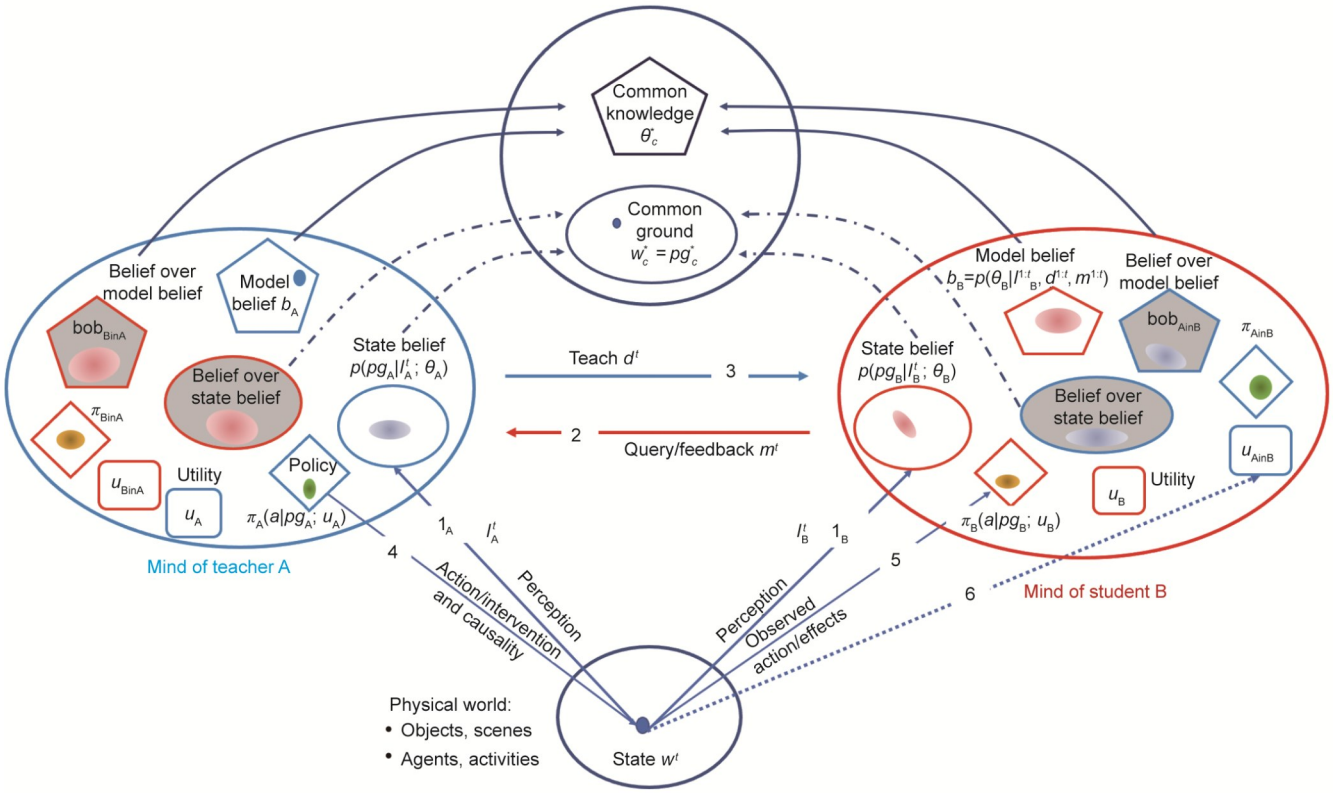

图7. 通学表征的放大图：统一了所有现有的学习协议并且可以用来提出更多的机器学习算法。每个思维包含四个空间：①五边形表示假设/模型空间 $\boldsymbol{\Theta}$；②椭圆表示状态空间 W 中的 $w=pg$，其中信念由分布图表示；③菱形表示策略 $\boldsymbol{\pi}$；④圆角矩形表示效用 u。箭头说明了动态：观察、干预和信息。阴影的形状代表了信念的信念。下脚标用于表示老师（A）、学生（B）和他们的共同知识（C）。对于递归下标，组件的所有者出现在 in 后面。例如，π_{BinA} 是在老师 A 的想法里。图中的参数和变量在第 3.1 和 3.2 节中定义。

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

划，例如，以T-pg的形式。请注意，这个动作a指的是在执行任务中与世界交互的动作，而不是指智能体如何在通学模式中发送消息。后者是一种通讯协议，我们将在第4.3节详细说明。π_A和π_B表示老师和学生的政策。在实际操作中，为了表示这种从状态到动作的映射，我们通常假设策略函数遵循一定的表达形式，并且只考虑其参数。此外，策略函数通常是由智能体的价值决定的，这种价值通常被定义为效用函数。

（5）效用函数u: $W\mapsto\mathbb{R}$表示智能体的价值观，即智能体关心的东西、错误的损失以及动作的成本。与模型相结合，即世界运行的规律，效用函数可用于任务规划[71]。u_A和u_B表示两个智能体的效用函数。通学中的智能体A和B还必须估计和学习其他智能体的效用函数，我们分别用u_{BinA}、u_{AinB}表示。我们将在后面讨论u_{BinA}、u_{AinB}会影响通学协议和学习过程的平衡，导致不同的学习极限。

3.2. 统一的学习框架

为了总结上一节的内容，我们在图7中汇总了通学模式所涉及的表征。总共有六套思维。每个智能体都有两套不同颜色的思维，一个代表自己，一个代表其伙伴。这四个思维中的每一个都有四个组成部分（w, θ, π, u），由不同的形状代表。智能体通过保持对w和θ的信念和信念的信念来刻画它们的不确定性。另外两套思维描述了共识和外部世界，即上帝的思维。

图7中的箭头显示了各种动态和信息流，包括三个类型。

- 观察I_A和I_B从物理状态到感知状态空间W。
- 引起物理世界变化的变化的动作或干预。
- 两个智能体之间交换信息的消息。根据学习模式的不同，这些信息用于推理、学习、演示、确认等。

为了清楚起见，我们省略了其他动态的箭头：例如，有些信息可能是由信念的信念空间产生的，以探测其他智能体的想法，如“我认为你的状态w是……”或“你对状态w了解多少?”。有些箭头是二阶的，例如，老师通过观察学生如何进行任务来了解他的策略π_{BinA}，即通过示范学习（learning from demonstration, LfD）[30,75]，或者通过观察学生的决定或选择来学习其效用u_{BinA} [71]。

在通学中，老师和学生的交流在三个层面上趋于一致（见图7中弯曲的箭头）：

- 当推理过程收敛时，他们会达到对当前的情况达成共识w_c^*；
- 当学习过程收敛时，他们达到对模型达成共识θ_c^*；
- 当他们的政策和效用趋于一致时，他们会达成一个共同的社会规范π_c^*和道德即价值观u_c^*。

根据智能体的学习协议和特征（即产生和解释信息的能力），收敛可能有不同的均衡，决定了学习的极限。在通学中，我们假设智能体是合作的，没有欺骗性，而且他们的效用函数通过学习会达到一致。在大多数学习设置中，收敛是在第二和第三层次考虑的，我们将在本文的大部分内容中使用它们作为收敛标准。我们将在5.2节中讨论多层次的收敛问题。

我们提出通学模型的目的是将人类教育学的精髓融入机器学习，并克服现有算法的限制。事实上，根据目前的讨论，我们可以看出，通学模式实现了对于现有的机器学习方法的统一，其关系如下。

- 图3中的香农通信描述两个智能体之间的信息传递通道。通学通过对智能体思维状态、信念以及信念的信念空间、效用函数和共识的刻画，扩展了这种交流环境。另外这些成分都是可以随着时间而变化的，使得可以用更复杂的信息进行通讯，并让智能体能够“理解弦外之音”（read between lines）。
- 概率统计学习理论，是一种被动地从随机抽样的例子中学习的过程。可以是有监督或者无监督的。这由图7中的箭头1_B显示。与此相反，在通学中，信息是在反映思维状态和效用函数的基础上进行发送的而非随机生成的。
- 主动学习由箭头2_B表示，学生可以要求老师为他选择的某些例子贴上标签，选择这个例子是为了在优化B的效用/损失函数时获得最多的信息。
- 算法式教学[77–78]如箭头3所示，是和主动学习相对应的形式。老师选择最好的例子来教学生，以提高其效率。A必须考虑B知道什么，并选择关键的例子给B，如支持向量。
- 从示范中学习[15,30,75]是机器人学中的一个典型的学习范式，是常识获取的一个重要组成部分。这种学习方法由图7中的箭头4和5所示。老师通过对物体的一系列动作来传授一项任务。学生直接观察这些动作和它们的结果，并从老师那里学习动作策略。
- 因果学习用箭头1_B和5表示，智能体应用动作来改变物体和场景的要素，并学习其动作在改变物体要素方面的因果关系，包括外观变化（如刷墙、拖地）、几何变化（如吹气球）和拓扑结构变化（如切水果）。

通学还可以创造新的学习方法或协议，例如：

- 感知因果学习。因果学习[79]要求智能体A做出动作、实验或者干预（箭头4），并观察其效果（箭头1_B和5）。在参考文献[73]中，我们提出了和因果学习不同的学

习方式——感知因果学习。其中，学生可以通过观察（箭头1_B和5）老师的动作（箭头4）来学习因果关系。在假设她没有施展魔法（即没有作弊）的情况下，学生会推断并反思老师的动作。这被称为“感知因果”。如参考文献[73]中所说，这种能力是人类智能关键，对学习因果关系十分有效，相关研究为人工智能从观察中学习因果关系打开了大门。

• 效用学习由箭头4和6显示。学生通过观察老师的决定和动作选择，推断出她的效用函数。经济学理论认为，理性的智能体做出决定和采取动作是为了实现效用最大化。通过观察A采取的动作，B可以推断出A的效用，在通学中用u_{AinB}表示。例如，在参考文献[71]中，我们展示了人工智能折叠T恤衫，通过观察老师的折叠动作，学生不仅可以学到因果关系和策略π_A，还可以学到美学上的效用函数u_{AinB}：T恤的哪些状态对老师来说具有相对高的价值。B可以选择一个相近的价值$u_B \leftarrow u_{\text{AinB}}$。在通学中，智能体将更新和调整到一个共同的效用。我们将在后面的4.3节中展示另一个价值对齐的例子。

• 类比学习（learning by analogy, LbA）是人类使用的一种强大的学习模式[4]，但在目前流行的机器学习方法中通常是缺失的。它要求两个智能体之间有一定共识（w_c, θ_c），并要求对知识有抽象和投射的能力，以使用图形表征进行跨领域的知识转移。抽象和投射是经典的瑞文智商测试（Raven's IQ test）中的关键能力，但在目前的统计学习中却没有涵盖。共识将促进类比学习。随着共识的累积，两个智能体将越来越同步，学生将逐渐变得和老师有相近的认知能力。

总而言之，上述学习范式都有其最合适的使用场景，说明了通学框架在各种特殊情况的使用。然而，真正的类人人工智能的部署通常涉及更复杂和全面的设置，所以完整的通学框架中每个组件都是必不可少的。在4.3节中，我们展示了一个具有足够复杂度的示例场景，以展示完整通学模式。现在，我们结束了对通学表征的介绍，接下来我们介绍通学模式中思维更新的具体形式以及学习是如何开展的。

3.3. 通讯式学习的模式

在3.1节中，我们介绍了通学模式的思维表征，有了它，人类教育学和心智理论与机器学习的整合成为可能。在本节中，我们将展示老师和学生的思维动态。这完善了通学模式的定义，使我们能够通过实例化这个一般化的模式来审视已有的机器学习算法。

3.3.1. 整体设置

首先，让我们回顾标准的机器学习算法。一个算法是从训练数据到模型空间的映射[18,40]，其中模型可以是，如支持向量机中的一个特定的超平面、*K*-means中的K中心点的位置或神经网络的参数。在通学中，我们将学习概括为一个教育学过程，涉及两个智能体，一个老师和一个学生。每个智能体都有自己的模型，或者在需要考虑不确定性的情况下，有一个模型的信念。在3.1节中，我们区分了智能体的模型、策略和效用效用函数，以更好地说明在情景交流场景中的各种学习类型。然而，对于学习动态的建模，我们不需要明确地对它们进行区分。也就是说，我们在这里提到的模型是一个具有更广泛含义的表示，而不仅是3.1节中的模型，后者的唯一目的是为了解释物理世界。在后面的实例（见3.4节和附录A中的S3节）中，我们将看到概念、策略、价值和效用都可以作为具有相同通学模式的模型，共享同一个学习的动态函数。

让我们把老师的模型空间和模型[†]的信念表示为Θ和$b_\theta \in \Delta(\theta)$[†]。我们假设$b_\theta$在整个教学过程中保持不变，因为老师不会收到任何新的信息来帮助她更新模型。此外，在大多数情况下，我们假设老师知道真正的模型θ^*，也就是说，$b_\theta(\theta)$变成了$1(\theta=\theta^*)$。同样地，我们有学生的模型空间Ω和学生对模型的信念$b_\omega \in \Delta(\Omega)$。请注意，我们并不假设老师和学生有相同的模型空间，即可能Ω与Θ相同，也可能不相同。有两个独立的智能体模型空间，通学就可以处理老师和学生对于世界有不同的表征的情况。例如，假设模型是一个从机器人的相机输入到动作命令映射的策略网络。那么，具有不同摄像头配置的机器人仍然能够相互教学和学习。由于学生将更新他对模型的信念，我们用b^t_ω来表示他在时间t的信念。同样的上标将适用于其他随时间变化的变量。

通学的目标是让学生对模型有足够准确的信念，这样他就能在特定的任务中与老师达到相当的表现。我们可以把这个指标定义为最小化损失函数$L(b_\theta, b^T_\omega)$，其中，T是学习过程终止的时刻。L衡量的是老师的表现和学生的表现之间的差距。成绩差距越小，L越小。这里，为了确定通学模式，我们不需要具体定义L。在附录A的S3节中，我们将看到在不同领域通学模式的应用，并进一步具体化L。

[†] 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.

3.3.2. 老师的设置

通学是一个教育过程，在这个过程中，学生通过来自老师的信息更新自己对模型的信念。这里我们把老师在时间t的信息表示为$d^t\in\mathcal{D}$，从她的信息空间$\mathcal{D}$中选择。每个时间的信息选择标准取决于学生在那一刻的学习状态，表示为$s^t\in S$。学习状态是一个由老师维护的变量，用于跟踪学生在特定时间的进展。例如，s^t可以是时间t时学生的验证错误（validation error）。有时老师无法知道确切的学习状态，那么她需要对学习状态有一个信念$b_s\in\Delta(S)$。就像L一样，在研究特定的学习范式之前，我们不需要给$\mathcal{D}$和S的表示做出额外限制。

一般来说，当老师对学生的学习状态有一个准确的估计时，教学可以更加有效。因此，老师有一个学习状态的过渡模型。也就是说，学生在收到信息后将如何取得进展。假设是学习状态遵循马尔科夫假设（Markovian），过渡模型在数学上可以定义为：

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

也就是说，ψ的输入是学生当前的学习状态和老师的信息，并返回学生的新学习状态的分布。在有些情况下，比如主动学习，老师也接收来自学生的信息（查询数据）。我们可以用$m^t\in\mathcal{M}$来表示学生在时间t的信息。在不失一般性的情况下，我们假设在每个时间t，老师先发送d^t之后学生发送m^t。来自学生的消息也可以帮助老师估计他当前的学习状态。所以，老师对学生如何发信息也有一个模型：

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

将学习状态映射到学生的信息分布上。有了ψ和ϕ，老师可以用贝叶斯滤波器更新她的b_s，即

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

一般来说，这个信念不能被精确计算，特别是当$|S|$很大甚至无限大的时候。然而，在实际操作中，ψ和ϕ通常被建模为指示函数（indicator function），从而有效地简化计算过程。有了学生当前的学习状态和她对模型的信念b_θ，我们就可以为老师制定教学策略：

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

其中，$\mathrm{softmax}_\beta(x)=\exp(\beta x)/\sum_{x'\in X}\exp(\beta x')$是玻尔兹曼理性函数（Boltzmann rationality）[80–82]，而$Q_\psi(b_\theta,b_s^{t-1},d^t)$是一个价值函数，输入是老师对当前学习状态的信念、老师的模型和老师的信息。通常情况下，它可以进一步扩展为简单形式的$Q_\psi(\theta,s,d)$，由$b_\theta(\theta)$和$b_s^{t-1}(s)$加权。由于这个函数可能取决于老师的学习状态的转换模型，它包含参数ψ。Q可以从数据中学习，也可以由专家来定义[12,83]。在附录A中的S3节中，我们将看到这两种方式的例子。

3.3.3. 学生的设置

在这一节中，我们定义通学模式中的学生。与标准的机器学习算法从数据到模型的映射不同，通学中的学生知道自己是从一个合作的老师那里学习。也就是说，学生知道他从老师那里得到的是选定的信息而不是随机的例子。为了给有老师意识的学生建模，我们从一个没有老师意识的学生开始，他的信念更新规则（即给定一个新的信息d^t）和常规的机器学习算法一样：

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

其中，$\pi:\Omega\mapsto\Delta(\mathcal{D})$是学生认为的教学似然函数（likelihood function）。一般来说这和老师真正采用的教学方法不一致。然而，正如我们在接下来的章节中所看到的，大多数情况下，当学生对这个函数有一个合理的近似时，学习是有效的。

接下来，对于一个有老师意识的学生，信息的选择不仅仅依靠老师的模型。利用所有可用的信息，一个有老师意识的学生认为的教学似然函数是：

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

然而在π中使用整个历史有一个缺点，那就是其复杂度会随着时间呈指数增长。我们知道，老师在教学时依靠两样东西：一个是她的模型θ^*；另一个是她对学生学习状态的估计s。因此，我们将$\hat{s}\in\hat{S}$定义为学生对老师心中自己学习状态的估计。然后，完整的历史$d^{1:t-1}$、$m^{1:t-1}$可以浓缩为b_s^{t-1}，公式（10）变成：

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

其中，$\pi:\Omega\times\Delta(\hat{S})\mapsto\Delta(\mathcal{D})$是考虑了老师心智理论的教学似然函数。也就是说，在每一时刻，学生都保持两个信念，一个是对模型的信念，一个是对老师对他印象的估计。从理论上讲，老师和学生之间的相互推理嵌套可以是无限递归的[62]。为了避免无法计算，在通学中，我们只建模有老师意识的学生，不进一步递归。为了更新$b_{\hat{S}}$，学生还需要两个函数来模拟$\hat{S}$在两类信息之后的变化。我们定义了

$$\zeta:\hat{S}\times\mathcal{D}\mapsto\Delta(\hat{S}) \tag{12}$$

$$\xi:\hat{S}\times\mathcal{M}\mapsto\Delta(\hat{S}) \tag{13}$$

作为学生对$\hat{s}$的过渡函数，即在老师发送（ζ）和重新接收（ξ）信息后，老师对他的印象将如何改变。学生的对应函数ϕ应该是从$\hat{s}\times\Omega$到$\Delta(\mathcal{D})$的函数映射，作为他对老师在学习状态为$\hat{s}$时给定模型的教学方式的估计。这可以通过π实现，因为任何一个$\hat{s}'$都可以写成Dirac-delta分布，即$\delta_{\hat{s}}(\hat{s}')=1\ (\hat{s}=\hat{s}')$。请注意，$\zeta$、$\xi$和$\pi$都是老师在学生心

中的心理变化的近似值，所以用Ω和$\hat{S}$代替Θ和S。

现在，我们可以写出$b^t_{\hat{s}}$的信念更新函数。不同于公式（7），我们希望在学生收到d^t后和m^t被送出前有一个中间变量，其目的是确定送出的最佳m^t。让我们用如下公式表示：

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

d^tZ作为学生在收到d^t后对老师对自己学习状态的估计的信念，其中，Z是一个泛化系数（normalizing factor），详细的推导见附录A中的第S1节。利用$\tilde{b}^t_{\hat{s}}$，学生可以定义给老师发信息的策略，即

$$p\left(m^t|\tilde{b}^t_{\hat{s}},\, b^t_{\omega}\right) = \mathrm{softmax}_k\left(V_{\xi}\left(\tilde{b}^t_{\hat{s}},\, b^t_{\omega},\, m^t\right)\right) \quad (15)$$

其中，V_{ξ}（$\tilde{b}^t_{\hat{s}}, b^t_{\omega}, m^t$）是学生的一个价值函数。这里$V$以信念为参数，因为学生在选择信息时，通常需要考虑分布的属性，如熵（entropy）等。就像老师的价值函数Q一样，V也可以从数据中学习或定义。在学生向老师提供了他的信息m^t后，他用以下方法完成对于$\hat{s}$的信念更新：

$$b^t_{\hat{s}}\left(\hat{s}\right) = p\left(\hat{s}|d^{1:t}, m^{1:t}\right) \propto \int_{\hat{s}'\in\hat{S}} \xi\left(\hat{s}|\hat{s}', m^t\right)\tilde{b}^t_{\hat{s}}(\hat{s}')\mathrm{d}\hat{s}' \quad (16)$$

式（16）结束了通学中所有的信念更新函数。由于它涉及许多思维成分和智能体之间的嵌套推理，我们用图8来总结通学框架，并在算法1中给出了具体的学习步骤。在下一节中，我们将展示各种学习范式如何可以表达为通学的特例。

3.4. 用通讯式学习刻画已有的学习范式

第3.3节中定义的通学模式给了我们一个统一的视角来总结现有的学习范式。也就是说，我们可以通过构建相应的价值和思维更新函数，用不同的学习范式来实例化通

算法1 通讯式学习

Input-Teacher: $\Theta,\ b_\theta, b^0_{\hat{s}}, Q,\ \psi, \phi, \mathcal{D}$

Input-Student: $\Omega, b^0_{\omega}, b^0_{\hat{s}}, V, \zeta, \xi, \pi, \mathcal{M}$

Input-World: $T,\ L$

Output: b^T_{ω}

For $t = 1{:}T$ **do**
- Teacher selects d^t using Eq. (8)
- Student updates $\widetilde{b}^t_{\hat{s}}$ and b^t_{ω} using Eqs. (14) and (11)
- Student selects m^t using Eq. (15)
- Teacher updates b^t_s using Eq. (7)
- Student updates $b^t_{\hat{s}}$ using Eq. (16)

End

Return $b^T_{\omega}, L(b_\theta, b^T_{\omega})$

学模式。由于并非所有机器学习范式都包括通学中全部的递归相互推理，我们可以根据它们的递归水平进行分类。如果一个学习范式只有一个学生，数据来自一个随机的过程，而不是一个合作的老师，我们称它为0级范式。一些0级范式可以涉及多种智能体，如联合培训/教学[84–86]的标签噪声学习的框架。尽管有两个学习网络，但它仍然遵循被动学习的变化，因为学习智能体之间没有相互推理，它们只在参数初始化方面有所不同，并在训练过程中寻求保持这种差异，以避免确认偏差。对于像主动学习[29]这样的范式，其中老师只是被动地回答学生对于随机生成数据的询问（query）而没有自主性，我们也将其归类为0级。假设有一个合作的老师可以选择她给学生的信息，但学生不知道这个老师的存在，这样的学习范式被称为第1级。当老师在信息选择方面是合作的，同时学生也意识到她的帮助，我们就有了第2级范式。

我们在表1中总结了这些范式和它们的级别，可以看出，所有这些学习范式的本质都是通学的一个或几个成分被省略得到的特殊情况。在每个学习范式中，可以有各种

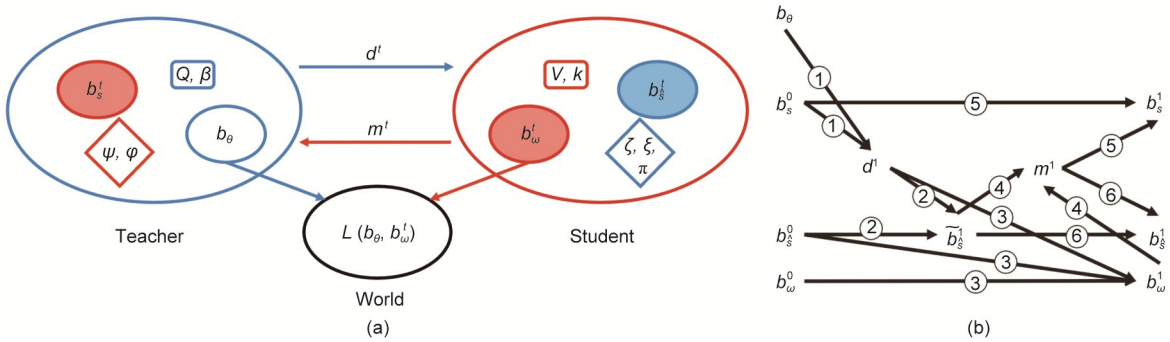

图8. 通学模式。（a）老师和学生的心理状态表示。蓝色的气泡是老师的，红色的气泡是学生的。椭圆表示信念，其中阴影部分是随时间变化的信念。菱形表示智能体用来更新他们对伙伴的行为和知识信念的函数。我们称它们为信念更新函数。老师和学生的价值函数分别在蓝色和红色的圆角矩形中。（b）扩展了第一个时间步骤的过渡过程。图中的数字将箭头分组，相同的数字表示操作发生在一个函数中。箭头“1”对应于公式（8），箭头“2”对应的是公式（14），箭头“3”对应的是公式（11），箭头“4”对应的是公式（15）。公式（7）和公式（16）分别由箭头“5”和“6”表示。这些信念更新的时间顺序与箭头的数字顺序相同，其中，“2”和“3”同时发生，“5”和“6”同时发生。

各样的学习算法，每个算法都映射到CL组件的基础。在附录A的S3部分中，我们为每个范例选择了一些算法，并展示了它们的CL基础。由于本文的目的是提出一个统一的学习框架，而不是详尽地列出所有的学习算法，因此我们只为每个范式选择一些范例算法，并指定它们的CL基础。对于其他算法，它们的CL基础可以很容易地从属于同一范例的范例算法中迁移出来。这些基础的完整总结和比较见附录A中的表S2和表S3。

关于通学模式的一个后续问题是如何获得信念更新和价值函数。通常情况下有两种方法。第一种是使用预先设计好的、专门为特定的任务而设计的启发式函数（heuristic）。4.3节中的人机交互（human-robot interaction, HRI）合作任务就属于这一类。第二种是通过师生不断地配合，学习出适合的相应函数。4.1节中的指代游戏（referential game）的例子属于这一类。也就是说，在学习和传授多个模型后（每个模型的学习都遵循算法1），老师和学生开始更好地了解对方，并形成一种默契的通讯规范（communication norm）。现在让我们通过一些具体案例展示通学模型的应用。通过这些例子，我们将发现信念更新和价值函数是如何定义或学习的。

表1 之前工作中的学习范式及其递归水平

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. 通讯式学习的实例

到目前为止，我们已经完成了对通学模式的定义。本节将通过几个例子来证明该模式的可行性和必要性。首先，我们展示了一个具有语用学（pragmatic）协议的有效性，该协议是通过将通学模式应用于指代游戏而产生的[87]。然后，我们介绍通学在一个现实的人机交互任务中的应用。

4.1. 实例研究1——在指代游戏中浮现出包含语用学的学习协议

在本节中，我们用一个概念验证的指代游戏作为例子来证明通学模式的使用。尽管这个游戏很简单，但它包含了协作智能体之间标准通信的必要组成部分。已经有研究证明，在玩各种形式的指代游戏的过程中，智能体之间可以出现有效的通信协议[88–89]，但在这个例子中，我们阐述了通学模式是如何促进智能体间出现具有语用学的通信协议。此外，我们还展示了在3.3节中定义的信念更新和价值函数是如何通过合作性的相互交流，以指代游戏的结果为监督从头学出来的。

4.1.1. 背景——指代游戏

我们在2.3节中提到的例子实际上是一个指代游戏。在一个指代游戏中，有一个老师、一个学生和一系列物品。老师心中有一个目标物，她需要向学生发送一个信息，使学生在收到这个信息后能够从这些物品中识别出目标物。之前介绍的通学的基础结构使通过心智理论建立适当的通讯成为可能，即老师必须考虑到学生此时的思维，而学生同时也必须在理解信息时考虑老师为什么这么说，而不只是理解信息的字面意思[83]。在图9中我们展示了一个例子。有三个物体：一个蓝色球体、一个红色球体和一个蓝色圆锥体。假设目标是蓝色球体。如果信息只可以是颜色或形状，那么，对于一个只理解字面意思的学生来说，蓝色球体没有唯一的标识符，因为“蓝色”和“球体”都有不止一个一致的物品。但是，可以意识到老师的合作性的学生，在听到老师说的“蓝色”后，应该能够进行语用推理，确定蓝色球体而不是蓝色圆锥体，因为他知道老师选择信息是有目的的而非随机的，所以会用“圆锥体”毫不含糊地指称蓝色圆锥体。在有语用学[10]的通讯中，信息不止传达字面意思。信息的用法通常还可以暗示老师的意图。

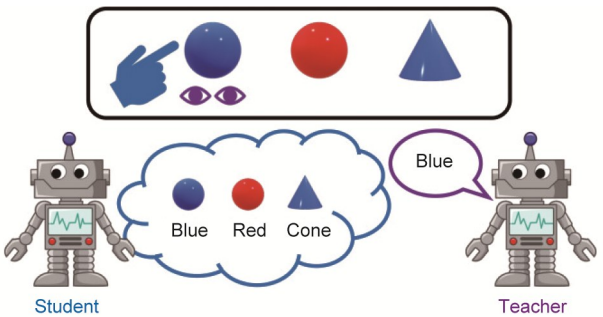

图9. 一个指代游戏的例子。有三个物体：一个蓝色球体、一个红色球体和一个蓝色圆锥体。如果学生听到老师说“蓝色”，他应该能够识别蓝色球体而不是蓝色圆锥体。

4.1.2. 包含语用学的学习协议的浮现

语用学研究的是语境（context）如何有助于语言的意义。在人类交流中，语言永远不会脱离上下文而被理解，句子通常可以传达比其字面意思更多的信息。然而，这种机制在大多数多智能体系统（multiagent system）中是缺失的，这限制了交流的效率和人机互动的能力。图9中的

例子具体说明了一个典型的语用学原则，即数量的含义（scalar implicature）。更多的细节将在5.1.2节中讨论。只有当智能体拥有心智理论并且能够捕捉到他们伙伴的合作意图时，他们才能进行语用学的推理。幸运的是，通学的基础设施支持心智理论，并允许出现语用学的协议。

4.1.2.1. 概述。在说话之前，老师会遍历所有的信息，用ψ预测学生在收到每个信息后的新信念。然后，她会发送可以产生最理想新信念的信息。听到这个信息后，学生会更新他的信念并采取动作（做出指代判断或再等待下一个消息）。心智理论中的智能体间的递归相互推理直接被整合到信念更新过程中。在我们的模型中，信念是老师定义在公式（3）中的Q函数的一个输入。而学生则会根据他对于目标物的信念直接采取行动。信念更新函数ψ和π的演化反映了智能体之间的通讯协议的变化。

4.1.2.2. 老师。老师根据她的Q值和信念更新函数ψ来选择信息，ψ的输入包括所有的物品、对学生当前的信念估计值和信息。这个函数的返回值是一个对新信念的估计。这个函数可以被参数化为一个神经网络，以softmax作为输出层。信念更新函数的返回值被直接输入Q函数。也就是说，信念更新函数的输出被用于Q函数，以预测学生在测试期间下一步的信念。老师根据公式（3）来选择信息。

要形成通讯协议，老师需要学习两个函数：ψ和Q。在训练阶段，每当学生收到一个信息，他就把他的新信念b_ω^t返回给老师。在测试阶段，老师则需要使用ψ的输出来近似预测学生的新信念。我们通过最小化b_ω^t和老师的预测之间的交叉熵（cross-entropy）来训练ψ。老师的Q是通过Q-learning [90]以指代游戏的结果作为奖励（reward）来训练的。

4.1.2.3. 学生。学生在指代游戏设置中，不需要给出太多的反馈。因此，我们只介绍他的信念更新函数π和决策策略。我们通过REINFORCE [91]算法直接学习π和学生的策略。在指代游戏中，学生的策略非常简单。如果他的信念b_ω^t足够确定，他就会根据他的信念指出目标；否则，就等待进一步的信息。π和策略可以被参数化为一个端到端的可训练的神经网络，输入是所有物品、之前的信念和收到的信息，并返回一个动作分布。

4.1.2.4. 适应性训练。老师和学生都接受适应性训练，以最大限度地提高他们的配合的预期成功率。我们先固定学生不懂，训练老师来更新她的Q和ψ。经过一段时间或直到性能收敛，我们固定老师并训练学生。如此交替训练直到师生配合成绩不再提高。到那时，有语用学的协议将最终浮现。

4.1.3. 协议分析

上一节中介绍的浮现出的通讯协议展示了有语用学的推理，它带来了比其他浮现的协议明显更好的指代成功率[87]。在图10中，我们展示了指代游戏的例子，以及训练期间老师和学生的信念变化。很明显，学生可以区分目标和干扰项，即使老师的信息与不止一个物品一致。也就是说，学生能正确地掌握信息的字面意义和语用学意义。具体来说，如果学生看到图10中的四个三维物体，他就会知道，老师会用“蓝色”指代迷惑项1，“大”指代迷惑项2，“紫色”指代迷惑项3。因此，“右上角”和“椭圆体”虽然与多个物体一致，但必须一定表示目标物。从数学上讲，浮现出的语用学协议近似地估计了概念空间的递归教学维度（recursive teaching dimension, RTD），它衡量了一对合作和理性的智能体之间概念学习所需的例子数量[87,92]。直观地说，在一个概念空间中，有一个最容易学习的概念子集，即在所有概念中具有最小的教学集（teaching set）的那些概念。我们可以先学习这些概念，然后把它们从概念空间中去掉。现在，对于剩余的概念空间，我们可以继续学习最简单的概念，以此类推。这种学习模式的教学复杂度一定不高于常规的教学维度（teaching dimension）[93]。在我们迭代训练的每一个阶段，智能体都会学习识别剩下的“最简单”的物品可以如何最优地指代。在我们的案例中，可以用唯一信息指代的物品是最简单的。

简而言之，指代游戏的例子说明了通学模式可以如何用来从头开始，只依靠通信结果的信号，发展出新的学习协议。接下来，我们给出了通学在一个比指代游戏更真实、更复杂的环境中的使用例子，以表明该模式的通用性和可扩展性。

4.2. 案例研究2——迭代式有老师意识学习

在参考游戏中，老师的目标是向学生指出干扰物之间的目标。这个过程可以被表述为一个具有离散概念空间的概念学习问题[18,94]。概念空间的可处理性甚至有限性使得参考博弈和类似的概念学习问题成为研究CL优势的合适起点[45,55]。然而，为了覆盖全部的机器学习范式，CL必须适应棘手的假设空间的问题，如学习连续参数[38–41]。

本节以机器参数学习为例来研究CL的理论收敛保证。比较了两种类型的学习模式，第一种是一个协作性的

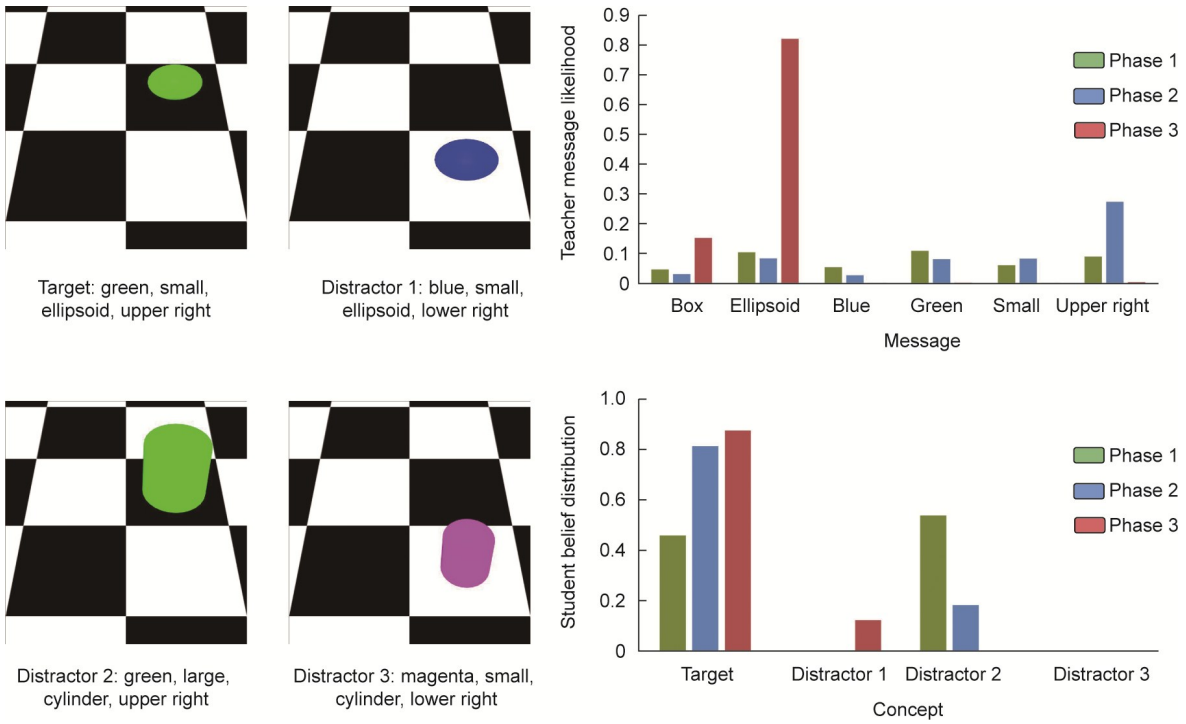

图10. 有4个物品的指代游戏。在每个阶段结束时，老师和学生的固定和训练角色都会互换。我们展示了用于指代目标物的信息分布和学生在收到最可能的信息后的新信念。老师用于指代迷惑项的信息分布，在第一阶段训练后所有的概率权重都集中到相应迷惑项所具有的唯一指代信息上。学生的信念说明，老师用于指代目标物的信息虽然与多个物品一致，但随着训练的进行，还是可以成功地指出目标，而且越来越有把握。一般来说，两个智能体的行为都会变得更加确定，而且确定的行为刚好可以配合得更好。

机器老师和一个无老师意识的学习者；第二种是同一个老师和一个有老师意识的学习者[95]。学习者用分布估计老师的数据选择过程，并用这个估计修正他的似然函数，以适应老师的意图。最大化新的可能性，使学习者能够利用从所选数据中获得的显式信息和在教学背景下提出的隐式信息。很明显，遵循CL形式主义的师生群体在经验上和理论上都能取得更好的表现。

4.2.1. 背景——机器参数教学

由于其连续的状态空间和长范围的规划，优化参数教学一直是一个具有挑战性的问题。到目前为止，最普遍和最有前景的框架是机器教学[39–40]。因此，我们采用了机器教学的迭代变化[41]。遵循3.3节中的符号，老师持有 $b_\theta(\theta) = 1(\theta=\theta^*)\theta^*$，其中，$\theta^*$是只有老师才知道的固定的地面真实参数，可以通过最小化定义在遵循分布的数据上的损失函数来获得，如回归的均方误差、分类的交叉熵损失和逆强化学习（IRL）的负对数似然[96–97]。在本节中，我们假设损失函数的选择是老师和学生的共同知识。

4.2.1.1. 概述。老师和学生可以用不同的但确定性相关的方式来表示相同的数据示例。那么，老师和学生的最佳参数也位于不同的空间。这模拟了实际场景：老师和学习者是一个人和一个机器人，或者是两个以不同方式制造的机器人。因此，按照3.3.1节中的符号，我们分别使用 $\theta^*\in\Theta$ 和 $\omega\in\Omega$ 表示老师和学生的参数。由于例子的表示可能很复杂，如DNN提取的特征[25,27]，在不损害表达性的情况下，我们假设最终的决定是通过应用线性模型来做出的。

为了避免老师和学习者之间无限的相互推理，限制了他们之间的相互知识。在本节中，考虑一个老师，他假设无老师意识的学习者正在使用随机梯度下降法（stochastic gradient descent, SGD）来更新他的模型。与此同时，学习者知道这些数据来自一个老师，而不是随机的。也就是说，无老师意识学习者、老师和具有老师意识学习者分别具有0级、1级和2级递归推理，如第3.4节所述。这些递归水平模仿了人类的认知能力[65–66,98]，也在参考文献[53]中被采用。

4.2.1.2. 老师按照机器教学的标准设置,老师和学生只能通过实例进行交流。这个限制并不影响框架的通用性,因为这些示例可以具有通用的格式,如在IRL中使用的演示[96–97,99–100]。如在3.3节中定义的,数据 d^t 是迭代提供的。老师的目标是为学生提供有用的例子,使他的参数 ω 尽快地收敛到最优值 ω^*,因为老师无法进入 ω^t 或者 ω^*,学生会在每次迭代中给出一些反馈 m^t,让他不断更新教学的进展。

4.2.1.3. 无老师意识的学习者。无老师意识的学习者使用简单的学习算法，例如，SGD [41–42,101–102]进行基于迭代梯度的优化。

4.2.2. 迭代的具有老师意识的学习者

我们首先描述了学生应该注意的老师，与参考游戏中的同伴相似，参数学习中的老师根据其Q值和置信更新函数来选择信息。在4.1.2节，我们描述了如何通过让老师和学生玩参考游戏来学习这些功能。由于较大的参数空间和较长的视野规划，老师使用贪婪启发式定义Q函数，即学生参数与真实参数之间距离减小最多的示例，这是机器教学问题中常用的标准[39–41,53]。该实例所带来的改进被定义为其教学量[41]，并作为其Q值，其数学定义可在在附录A的S3.2节中的等式（S40）中找到。因为老师在大多数实际设置中，没有访问学生的当前参数，即使他知道参数，也不能直接在不同的参数空间使用该值，我们让学生将其参数和数据的内积返回给老师作为反馈[41,95]。在线性模型假设下，老师可以使用简单的置信更新函数来捕捉学生的学习状态。详细信息请参见附录A的表S2中的迭代机器教学（IMT）[41]行。

IMT提出了一种协作性的机器老师，它大大帮助了初级学习者，但由于忽视了学生的老师意识，仍然缺乏完整的教学方法。为了填补这一空白，我们利用了CL，并提出了迭代老师感知学习（ITAL）[95]。有老师意识的学生通过考虑老师的例子选择来调整自己的π函数。直观地说，给定一个小批量中的所有数据，老师选择一个特定的例子，而不选择其他例子。用什么参数可以使这种选择的概率最大化？为此，更新后的教学似然函数有两个组成部分：第一个模型是参数和数据之间的一致性（示例的字面意义）；第二个模型是用估计的教学量和反事实推理（示例的实用[10]含义）。在新方法的帮助下，ITAL在各种回归、分类和IRL基准上实现了理论和经验上的改进[95]。图11 [95]比较了有老师意识的学习者和无老师意识的学习者的学习曲线。

ITAL不仅证明了CL的实用性，而且还提供了对通用的人机协作的见解，因为快速的参数学习使机器能够快速地适应用户的需求，甚至是实时的。在下一节中，我们将介绍CL在人机协作设置中的使用。

4.3. 实例研究3——双向人机价值对齐

通学模式的首要目的是实现类似人类的学习能力，从而实现通用的人机合作。机器可以通过人类用户的输入实时改变其行为，以便系统和人类用户合作完成一项共同的任务。要做到这一点，机器需要主动推断人类用户的信念、需求和目标[103–104]。自然地，这个推理过程可以被建模为一个学习问题，并被纳入通学模式。在这一节

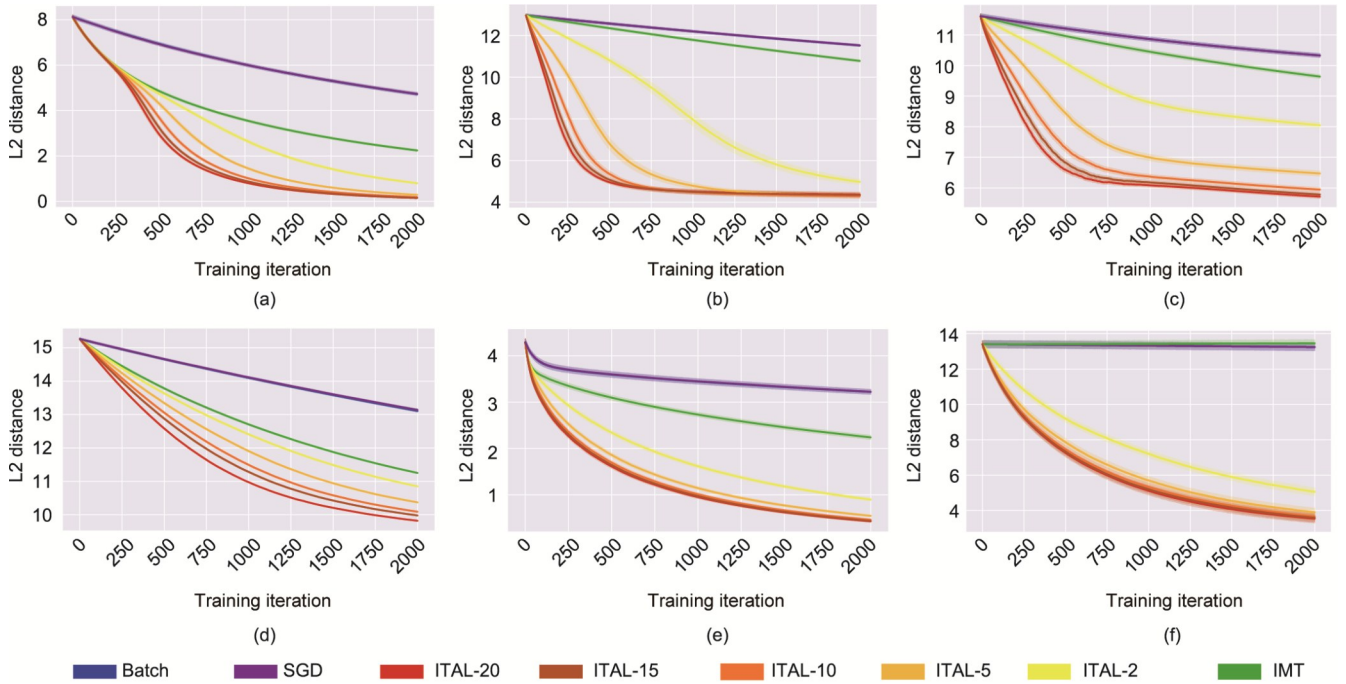

图11. ITAL在不同任务上的表现。（a）线性回归；（b）高斯数据；（c）CIFAR-10；（d）微型ImageNet；（e）方程；（f）在线IRL。在相同协作老师的情况下，ITAL总是比IMT取得显著的绩效提高，证明了CL带来的老师意识的好处。在2000步内，ITAL已经显示出收敛，而无老师意识的学习者在大多数任务中只学习有限的程度。L2距离的计算使用地面真实参数$\omega^* \in \Omega$和学生的当前参数。批处理和SGD用随机抽样数据教无老师意识的学习者。ITAL-X表示有老师意识的学生用来估计老师数量的小批量的大小。更多的细节可以在参考文献[95]中找到。

中，我们介绍了一个人机协作游戏，在这个游戏中，一队侦查机器人需要不断地将他们的价值与人类指挥官的价值对齐，以完成某些任务[105]。然后，我们展示了通学模式如何应用于这个场景。正如人们所看到的，游戏的成功完全依赖于指挥官和侦察机器人之间的有效沟通，这一点可以通过3.1节中的思维表征和3.3节中定义的通学模型得以实现。

4.3.1. 对于一般性人机合作的通讯范式

为了实现通用的人机合作，机器人必须能够了解用户的价值观，并实时改变自己的行为，以便人机团队能够合作实现一系列的共同目标。为了即时掌握用户的信息，传统的数据驱动（data-driven）的机器学习方法需要被团队合作中的通学模式所取代。这种以合作为导向的人机协作要求机器具备一定的心智理论水平：一台机器会主动推断出用户的信念、需求、目标[103–104]，以及人类的合作需求，因此形成一个以人为中心、与人类兼容的过程[106]。正如2.3节中例子所表明的，建立这种合作的本质在于共享的自主性（shared agency [107–108]）和共识（common mind [3,109–110]）。

4.3.2. 一般性人机合作的原型情境

我们设计了一个以协作游戏形式呈现的人机合作系统，在这个系统中，人类用户需要与一群侦察机器人合作完成一些任务并优化团队收益[105]。在这个游戏中，人类用户和侦查机器人在一个受限的频道（constrained channel）上进行通讯。只有机器人直接与物理世界互动；人类用户不能直接进入物理世界或直接控制侦察机器人的行为。同时，只有人类用户能够获得任务的真实价值函数（例如，使尽快地完成任务，收集更多的资源，探索更大的区域等）；机器人团队必须通过人机互动来推断这个价值函数。这样的设置真实地模仿了现实世界中的人机合作任务，因为许多系统都需要在人类用户的监督下，在危险的环境中自主运行。

要成功地完成游戏，机器人需要同时掌握“听”和“说”的能力来实现双向对齐。首先，机器人需要从人类的反馈中提取有用的信息，以推断出用户的价值函数并相应地调整他们的策略。其次，机器人需要根据他们当前的价值推断，有效地解释他们已经做了什么和计划做什么，以便用户知道机器人是否和人类有相同的价值函数。同时，人类用户的任务是指挥机器人侦察员到达目的地，并且使团队的得分最大化。因此，人类用户对机器人的评价也是一个双向的过程。人类用户必须推断出侦察机器人的目标，检查它是否与任务的给定价值函数相一致，如果不一致，则选择适当的指令来调整他们的目标。最终，如果系统运行良好，侦察机器人的价值函数应该与人类用户的价值函数保持一致，并且人类用户应该高度信任系统。图12说明了游戏中的双向价值排列过程[105]。在互动过程中，有三个价值：

• U_{A}：用户的真实价值；

• U_{AinB}：机器人对用户价值的估计，在这个游戏中，侦察机器人没有自己的价值，所以他们会根据 U_{AinB} 来行动；

• U_{BinA}：用户对机器人价值的估计。这是用户所具有的心智理论结构，对反馈和信任的形成至关重要。在这三个值中，发生了两个排列组合：

• $U_{\mathrm{BinA}} \rightarrow U_{\mathrm{A}}$：机器人从反馈中学习用户的价值；

• $U_{\mathrm{AinB}} \rightarrow U_{\mathrm{BinA}}$：用户从解释和互动中了解机器人的价值。

最终，三种价值将汇聚成 U_{A}，此时，人机团队将形成相互信任和有效的协作。

我们的设计鼓励人机合作和双向推理，因为双方在游戏开始时都拥有关键但只有自己知道的信息。侦察机器人拥有关于地图的信息，但无法获得人类用户的价值函数，而人类用户的价值函数决定了任务目标，没有它机器人无法做出符合人类用户意图的正确的决定。同时，人类用户知道支配决策过程的任务价值函数，但他无法直接访问环境。通过受限通信来实现人机协作，侦察机器人可以向人类用户发出行动提议，而人类用户则提供接受/拒绝的反馈，然后侦察机器人将利用这些反馈来推断人类用户的价值函数并相应地调整自己的行为。

从通学模式来看这个任务，我们把人类用户看作是老师A，把侦察机器人看作是学生B。学习的目标是让机器人的价值函数与用户的一致，并且让用户信任机器人，即 $u_{\mathrm{A}} = u_{\mathrm{B}}$，$u_{\mathrm{BinA}} = u_{\mathrm{A}}$。虽然在游戏中不涉及人际之间的模型学习，但同样的价值对齐算法只需稍作调整就可以应用于模型对齐。m^t 是机器人的提议和附带的解释，而 d^t 是用户对机器人团队的反馈（接受或拒绝）。在我们的设定中，通讯的主要目的是调整人类用户和侦察机器人之间的价值函数。在通学模式，人类用户是老师，试图通过通讯将她的价值传授给机器人学生。为了快速对齐，学生们需要知道何时以及如何做出任务提议。这样，来自老师的反馈对于正确估计价值函数是最有参考价值的。这些反馈直接改变了机器人的信念。为了从人类老师那里获得指导性的反馈，机器人和人必须建立起共识，即人类用户知道和相信什么，打算做什么，以及什么是一致或不一致的。只有在

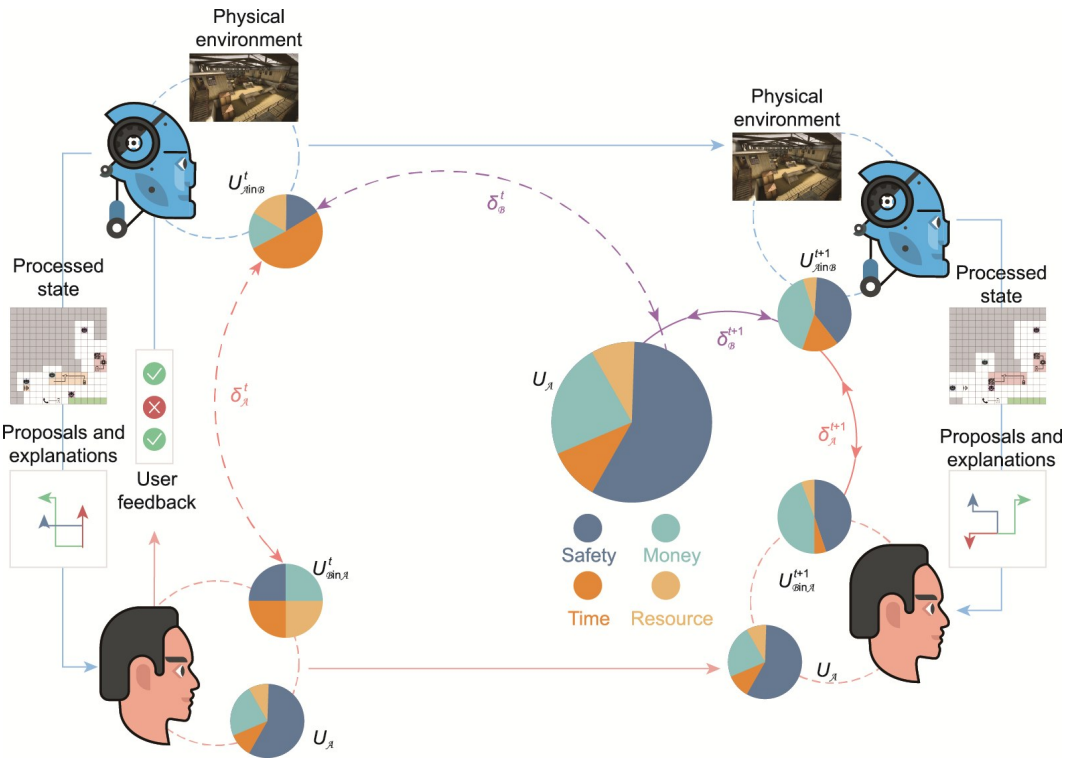

图12. 人类与机器人的双向价值对齐。饼状图表示价值，即不同目标在协作任务中的重要性，如同时考虑安全、获得金钱、节省时间和保留资源。上标中的t代表时间步骤。下标中的A和B分别代表“用户”和“机器”。U_{A}是用户的真实价值，U_{AinB}是机器人对用户价值的估计，U_{BinA}是用户对机器人当前价值的估计。δ表示任务价值空间中各价值之间的距离。在每一轮互动中，机器首先接收来自物理环境的信号，并处理其观察结果以形成环境的抽象状态。接下来，机器将处理过的地图连同任务的计划提议和解释一起呈现给人类用户，人类用户将向系统提供反馈，根据人类的价值和当前的地图状态接受/拒绝这些建议。考虑到用户的反馈，机器会更新其对人类价值的估计，并根据新的价值采取相应的动作。通过适当的解释，人类与机器人的合作交流使团队的价值在两个方向上保持一致，减少了U_{AinB}和U_{A}以及U_{BinA}和U_{AinB}之间的距离，从而最终收敛到真正的价值U_{A}。

这个共识的基础上，侦察机器人才能提供解释，说明以前的动作和当前的建议。这些解释会影响人类用户对于机器人信念的信念。

4.3.3. 游戏的设置

我们的合作游戏涉及一个人类指挥官和三个侦察机器人。游戏需要在一张未知的地图上找到一条从基地（位于地图的右下角）到目的地（位于地图的左上角）的安全路径。该地图被表示为一个部分可见的20×20网格图，每个格子都可能有一个不同的设备，只有在侦察机器人靠近它之后才可见。

我们为侦察机器人制定了在寻找到路径时额外的一系列目标，包括：①尽快到达目的地；②调查地图上的可疑设备；③探索更大的区域；④收集资源。游戏的表现是由侦察机器人完成这些目标的情况和它们的相对重要性（权重）来衡量的。其中的权重就是人类用户的价值函数。注意，这个价值函数只在游戏开始时向人类用户透露，而不是侦察机器人。游戏界面如图13所示，图14总结了人机互动的流程。

4.3.4. 用通讯式学习进行价值对齐

为了估计人类用户在通讯过程中的价值函数，我们将两个层次的心智理论整合到我们的计算模型中。第一级心智理论考虑合作性假设。也就是说，给定一个合作的人类用户，所接受的提议与被拒绝的提议相比，更有可能与正确的价值函数相一致。第二级心智理论进一步将用户的教育方法纳入模型。也就是说，促使机器人的价值更接近真实价值的反馈比其他反馈更容易被选择。建模人类用户的教育倾向（pedagogical inclination）需要更高一层的心智理论，因为这要求递归递给用户的机器人模型进行递归建模。结合这两个层次的心智理论，我们将人类的决策函数写成一个由价值函数参数化的分布，并开发出一种具有闭式（closed-form）参数更新的学习算法[105]。

值得注意的是，与我们的人机合作框架有可比性但不同的设置是逆强化学习（inverse reinforcement learning, IRL）[111]。然而，逆强化学习的目的是在一个被动的学习环境中，根据预先录制的，来自专家的演示（demonstration）来恢复底层的奖励函数（reward function）。正如我们将在表1中所总结说明的，标准逆强化学习是一个

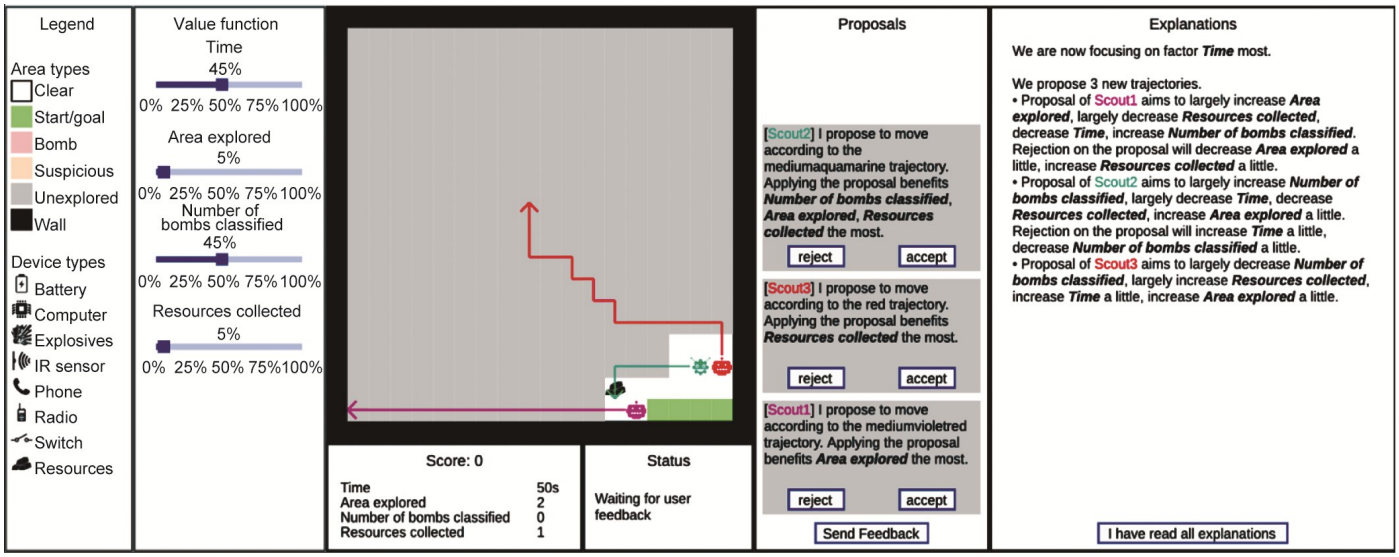

图13. 侦察探索游戏的用户界面。从左到右，图例面板显示游戏地图中的图例。价值函数面板显示这局游戏的价值函数，侦察员机器人不知道这个函数，用户也不能修改。中央地图显示当前地图上的信息。分数面板显示了用户的当前分数。总分是将各个目标的分数用价值函数加权后的总和。状态面板显示系统的当前状态。提议面板显示侦察机器人当前的任务计划提议，用户可以接受/拒绝每个建议。解释面板显示侦察机器人提供的解释。

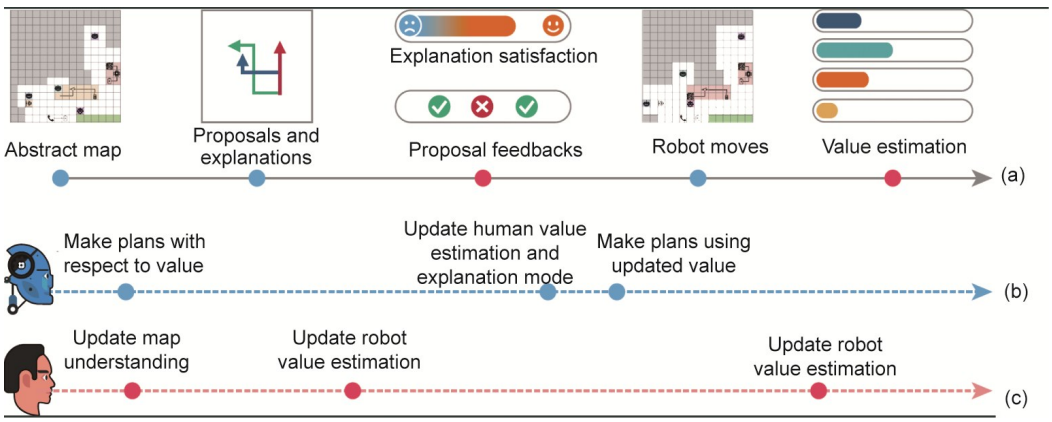

图14. 侦查探索游戏的研究设计。时间线（a）表示在一轮游戏中发生的事件，从机器人收到环境信号开始，到它们的下一步动作结束。根据用户的实验组，建议和解释以不同的方式呈现给用户。价值估计要求用户推断侦察机器人在当前的价值函数。这些问题的答案在游戏过程中不会被使用，而只是在游戏结束后用于测评用户的心理模型。时间线（b）和（c）分别描述了机器人和用户的思维动态。

0级范式。相反，在我们的环境中，侦察机器人被设计为从人类用户给出的稀缺监督中进行交互学习。更重要的是，我们的设计要求机器人在任务进行的过程中实时地主动推断人类用户的价值，这是属于第二级范式的以人为中心的学习方案的独特属性。此外，为了完成合作，侦察机器人不仅必须迅速理解人类用户的意图，而且还要阐明自己，以确保在整个游戏过程中与人类用户顺利沟通。简而言之，机器人的任务是通过推断人类用户的心理模型，积极提出建议，并评估人类用户的反馈来进行价值调整，这需要对人类用户进行复杂和反复的心理建模。

5. 对于机器学习基础理论的贡献

正如第2节所讨论的，目前在通信、应用数学和统计机器学习方面的框架仅限于特殊的设置，它们得出的性能界限如PAC-learning和Vapnik-Chervonenkis (VC)维度，往往过于悲观[77–78,112]。大多数概念在PAC-learning的设定中是“不可学习的”（not learnable）[9]，而人类智能可以从少量的例子中就可以学习大量的日常任务。主要原因是目前的方法没有考虑到人类交流的许多关键方面，最终导致学习协议的有效性降低。

在本节中，我们将讨论通学模式对于基础的机器学习理论的贡献。我们首先介绍一种新的学习表征，利用这个

表征，我们可以获得超越香农通信极限的学习协议[8]。然后，我们将讨论学习的层次，并在这些层次之上提出了学习的停机问题。

5.1. CL引入的学习表征

5.1.1. 当奥曼遇到格莱斯——从分布式知识到共识

自我们意识到香农通信模型和PAC-learning理论框架在刻画人类学习中的局限性后，我们寻求另一种通用的表征，足以同时涵盖科学家和学生的学习方式。正如我们在2.1节所讨论的这两种类型的学习都可以被解释为信息从一个思维传递到另一个思维的过程。因此，一个通用的表述应该同时模拟信息传递“出发地”的思维和“目的地”的思维。特别地，如果要模拟学习过程中老师和学生的思维变化，我们需要对学习过程中的已知和未知的内容有一个清晰的表述。幸运的是，利用知识论逻辑（epistemic logic），也就是知识的逻辑（the logic of knowledge）[116]，我们可以为知识和信念引入一个严格的数学定义。在20世纪70年代，罗伯特·奥曼（Robert Aumann）将知识分析扩展到多个智能体，并将共同知识（common knowledge）的概念应用于经济学和博弈理论[117]。共同知识再加上后来提出的分布式知识[118]，为表示智能体在学习前后的思维状态提供了一个理想的工具，从而让我们得以对整个学习过程进行建模。

知识建模的框架是以可能的世界（possible world）为基础的[61]。可能世界模型背后的直观想法是，有很多可能的世界状态。鉴于其当前的信息，智能体可能无法判断哪个可能的世界描述了现在的实际状态。如果一个事实（fact）在所有可能的世界中都是真实的（考虑到智能体的当前信息），那么就可以说智能体知道这个事实。具体来说，我们可以想象一个拥有模糊摄像头的机器人。由于模糊的相机接收到的视觉信号不够清晰，机器人无法区分每一个可能的世界，所以机器人需要保持一组世界，这些世界根据它所接收到的信号来看都是可能的。所以机器人如果知道一个事实，那么这个事实在所有这些可能的世界中都必须是真实的，否则机器人就会有怀疑而不确知这个事实。

当有两个智能体对世界进行推理时，我们就可以引入共同知识和分布式知识的概念。

• 共同知识：两个智能体都知道的事实，两个智能体都知道他们的伙伴知道，两个智能体都知道他们的伙伴知道他们知道，以此类推。

• 分布式知识：如果两个智能体充分结合他们的知识之后，他们会知道的事实。

换句话说，共同知识是两个智能体都知道的东西，他们都无法否认，而分布式知识在他们通讯并交换知识之前，任何一个智能体都可能不知道。因此，分布式知识总是至少和共同知识一样精确，通常比共同知识更精确。所以我们可以定义学习为让老师和学生的分布式知识成为他们的共同知识的过程。例如，有两个机器人，它们的摄像头都是模糊的，但是模糊的方式不同。当他们交流和分享他们的知识以进一步精确他们每个人可能世界集合时，学习就发生了。当它们的共同知识与它们的分布式知识相同时，学习就终止了，因为两个机器人都没有对方不知道的私人信息（private information）了。

我们在图15中展示了一个用知识论逻辑表示的学习过程的实例。假设Alice和Bob对世界的感知并不完美（就像那个带着模糊摄像头的机器人）。也就是说，他们不能观察到ω，而是只能观察到一些世界的投影：

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

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

其中，当ω在它右边时，每个输入$I_{\mathrm{A},i}, I_{\mathrm{B},i}\in\{0,1\}$等于1，否则等于0。它们的感知划分（partition）分别是$\prod_{\mathrm{A}}$和$\prod_{\mathrm{B}}$：

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

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

也就是说，当世界是$\omega\in\Omega$时，Alice无法区分$\prod_{\mathrm{A}}(\omega)$中的世界。换言之，她知道真实世界一定在$\prod_{\mathrm{A}}(\omega)$中，但她不知道它是$\prod_{\mathrm{A}}(\omega)$中的哪个元素。同样地，Bob知道$\prod_{\mathrm{B}}(\omega)$中的一个世界必须是真实的，但对$\prod_{\mathrm{B}}(\omega)$中到底哪个是确切的世界感到困惑。

然后，根据对划分的一些基础知识，我们可以确定Alice和Bob的共同和分布式知识。给定两个定义在集合S上的划分$\mathcal{P}$和$\mathcal{P}'$：

• $\mathcal{P}$比$\mathcal{P}'$更精细，如果$\mathcal{P}'\forall s\in S,\ \mathcal{P}(s)\subseteq\mathcal{P}'(s)$

• $\mathcal{P}$比$\mathcal{P}'$更粗糙，如果$\mathcal{P}'\forall s\in S,\ \mathcal{P}'(s)\subseteq\mathcal{P}(s)$

直观地说，如果划分$\mathcal{P}$比划分$\mathcal{P}'$更精细，那么$\mathcal{P}$给出的信息集（information set）不会比$\mathcal{P}'$所提供的信息集给出的信息量少（因为可能的状态越少，不确定性越小，具有的信息量越大）。两个划分的交（meet）由$\mathcal{P}\cap\mathcal{P}'$表示，指比$\mathcal{P}$和$\mathcal{P}'$都更粗糙的最精细的划分；两个划分的并（joint）$\mathcal{P}\cup\mathcal{P}'$，表示比这两个集合都精细的最粗糙的划分[61]。有了以上的定义，我们可以表述出共同知识和分布式知识。

如图15（a）所示，在Alice和Bob交流之前，两个分

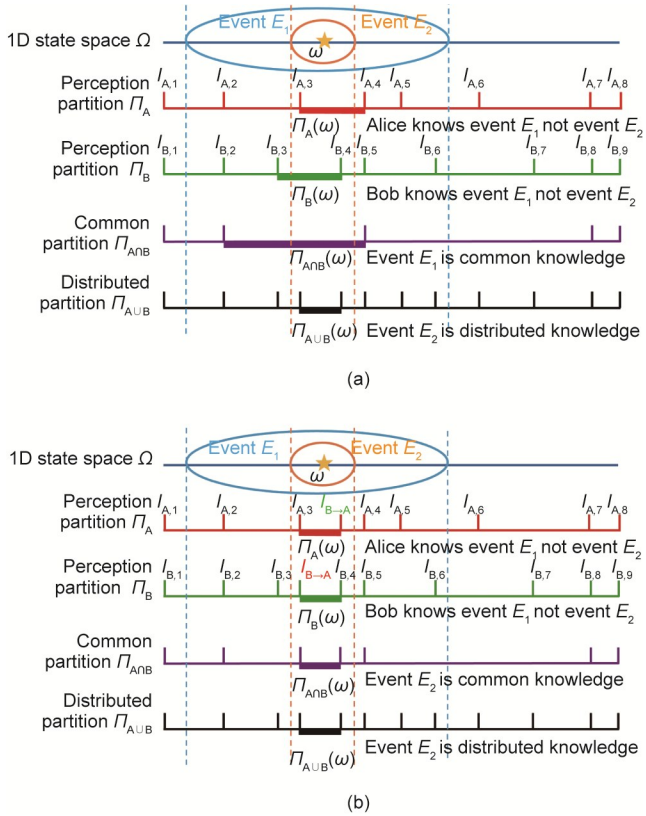

图 15. 用于推断 Alice 和 Bob 在一维空间中的状态 ω（星星）的共同和分布知识。每个线段代表划分（partition）中的一个块（cell）。落入同一线段的状态不能被区分开来。(a) 交流之前的知识表征；(b) 通信后的知识表征。Alice 和 Bob 的认知划分变得更加精细，因为伙伴的信息让世界进一步分化。Π_A 和 Π_B 表示认知划分；$\Pi_{A\cap B}$ 和 $\Pi_{A\cup B}$ 分别是两个划分 Π_A 和 Π_B 的相遇和连接。$I_{A,1}, \cdots I_{A,8}$ 和 $I_{B,1}, \cdots I_{B,8}$ 是观察的预测。

区的交 $\Pi_{A\cap B}$ 形成了 Alice 和 Bob 的共同分区，而两个分区的并 $\Pi_{A\cup B}$ 形成了他们的分布式分区。事件 E_1 的发生是众所周知的，因为

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

也就是说，由于事件 E_1 发生在 $\Pi_{A\cap B}(\omega)$ 的所有世界中，Alice 和 Bob 都知道它的发生，尽管他们不知道具体的世界。此外，Alice 和 Bob 都不能否认他们不知道 E_1，因为 $\Pi_{A\cap B}(\omega)$ 包含了 Alice 和 Bob 的所有认为可能的世界†，使 E_1 成为他们的共同知识。相反，Alice 和 Bob 都不知道事件 E_2 的发生，因为

$$\Pi_A(\omega) \not\subset E_2 \wedge \Pi_B(\omega) \not\subset E_2 \tag{22}$$

然而，通过结合他们的知识，E_2 也是普遍可知的，因为

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

在图 15（b）中，我们看到，通过将他们的一个分区边界分享给伙伴，Alice 和 Bob 可以将事件 E_2 从他们的分布式知识转移到他们的共同知识：

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

仅通过一轮信息传递，E_2 成为共同的知识，因为智能体认知划分的块（cell）被进一步压缩了：

$$\Pi_A(\omega) = \{\omega' : I_A(\omega') = I_A(\omega) \wedge I_A(\omega') = m_{B\to A}\} \tag{25}$$

$$\Pi_B(\omega) = \{\omega' : I_B(\omega') = I_B(\omega) \wedge I_B(\omega') = m_{A\to B}\} \tag{26}$$

在这里，我们略微滥用了符号，用等号来表示与收到的信息一致的认知。简而言之，使用知识论逻辑分析的框架，学习被建模为智能体间的通讯和分享知识，使每个智能体的世界划分得以细化，最终让信息从分布式知识传递到共同知识。

5.1.2. 超越香农极限——由通讯式学习带来的更好的学习协议

共同和分布式知识的概念为学习提供了一个正式表征。然而，只是知识论本身，仍然无法对人类教育学中的合作关系进行建模。具体来说，从分布式知识到共同知识的信息传递回答的问题是“什么是学习?”，但它并没有涉及老师和学生应该如何发送和理解信息，即“什么是高效的学习协议?”。为了回答如何高效地学习这个问题，我们必须在奥曼的知识表述基础上加入语用学（pragmatics）。

正如我们在 4.1.2 节中提到的，语用学是研究语言使用的语境（context）如何影响意义的语言学分支[10,36,119]。回顾一下图 5、图 9 和图 10 中的例子，老师和学生组成了一个合作的团队，不仅老师发出信息的字面意思，而且她选择某些信息的行为本身也会促进学生的学习。学生可以根据他们对情景、语境和老师的了解，对话语的含义做出了非常敏感的推断[83]。一个著名的例子是保罗·格莱斯提出的数量的含义（scalar implicature）：当人们说“我喜欢喝热水”时，尽管“热水”在语义上与“烫”相近，但它的意思是“不烫”；否则人们会直接说“烫”[120–121]。这个简单的现象包含了人类的两个基本特征，即对话者之间的协作性，以及递归的心智理论。通过在通讯的框架中的对学习建模，通学模式可以满足这两个单边机器学习范式不可能满足的条件。

图 16 用一个例子说明了有语用学的学习协议的优势。将一个图片定义为状态 ω，Alice 和 Bob 分别有 N_A 和 N_B 个神经元作为他们的观察：

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

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

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

其中，h表示神经元，可以是指标函数或图片的ReLU投影，就是说：

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

其中，λ是神经元的权重。如图16所示，π_a被8个红色神经元所包围，而π_b被4个绿色神经元所包围。假设Alice知道一个事件$\omega\in\pi_a$并通过以下方式告诉Bob：

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

然后，Bob将把他的认知从$\prod_B(\omega)$重新调整为π_a，实现的信息增益为：

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

如在等式（1）中所定义的那样。也就是说，Bob对可能的世界的信念从$\prod_{\mathrm{B}}(\omega)$缩小到$\pi_a$。有趣的是，如果Bob有心智理论并将语用学纳入学习协议，他就可以读出Alice的言外之意：Alice本可以但却没有使用那4个绿色神经元发送更短的信息，那么她的想说的是ω不在π_b中而在π_a中。因此，Bob可以进一步调整他的信念，实现的信息增益为：

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

其中，π_a/π_b指的是在π_a中但不在π_b中的区域。

命题：有语用学的协议比香农的通信协议更有效，因为使用前者时Bob获得的信息比后者多：

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

有语用学的协议通过整合心智理论，超越了香农的信息极限。额外的信息增益是通过考虑其他智能体的思维带来的。Alice在考虑了Bob知道什么之后选择信息，而Bob则会考虑为什么Alice发送这条信息而不是其他可能的信息。这样的反事实（counterfactual）推理增加了通讯的效率。

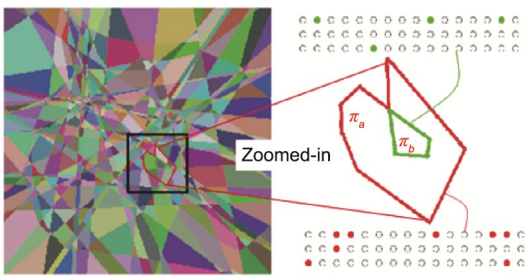

图16. 二维空间的划分和使用语用学协议推断状态的例子。左图：二维空间的划分。同一块内的状态会触发相同的神经反应。右图：π_a（红色）和π_b（绿色）所触发的神经反应。

5.2. 学习的停机问题

通过将学习建模成从分布式知识收敛到共同知识的过程，我们可以类比停机问题[122]提出“学习的停机问题”。也就是说，在什么条件下，学习过程会在不同的平衡点终止，这决定了学习的基本限度。就像我们日常生活中的传授和学习一样，通讯式学习是迭代进行的。对于迭代过程来说，都需要定义停止条件。到目前为止，在本文中，我们还没有深入探讨这个问题。在算法1中，我们用一个固定的步骤数作为停止条件。然而，事先确定适当的迭代次数是很困难的，如果不是完全不可能的。因此，我们需要一些标准来监控学习，并在学习达到极限时终止这一过程。

要解决学习的停机问题，我们必须知道学习的基本驱动力。也就是说，当老师和学生相互交流时，他们寻求达到什么共识。在这里，我们认识到学习的三个层次。

（1）信息层次：对单轮通讯中信息的理解。

（2）任务层次：老师和学生之间就某项特定的任务一致化彼此的思维，这个过程一般涉及多轮通讯。

（3）团队层次：了解伙伴的特征，并可以在不同的任务中重复使用。

每个层次都有一个独特的目标，并由通学模式中的一个循环控制，如图17所示，在每个循环中，老师和学生的目标是实现平衡。在下一节中，我们将详细介绍每个层次。由于通学的一个目的是促进新的学习范式的产生，我们在介绍的同时还包括了一些开放性的问题，以鼓励未来对相关主题的探索。

5.2.1. 学习的三个层次

5.2.1.1. 信息层次。信息层次表示老师和学生之间对每条信息的解释。由于信息是通讯过程的组成部分，能够完全理解说话者的意思是有效学习过程的先决条件。在这个层次上，信息会关于一个状态、划分中的一个块或一个事件（集合），以实现共识或共同的信念。尽管我们只讨论了几种类型的信息，如有标记的数据（3.4节）、作为划分的单元的投影（5.1.1节），以及线性神经元（5.1.2节），通学模式的信息空间可以扩展到解析图中的节点或逻辑表达。通学中的反思循环（reflection loop）负责信息层级的平衡。反思过程中涉及循环，是因为智能体具有心智理论并会进行递归的相互推理。特别是，老师考虑学生需要知道什么。而学生则思考为什么老师要发送一个特定的信息而不是其他信息等。因此，如图17所示，智能体形成了反思循环，包含了他们的自我信念b_θ/b_ω和他们的心智理论

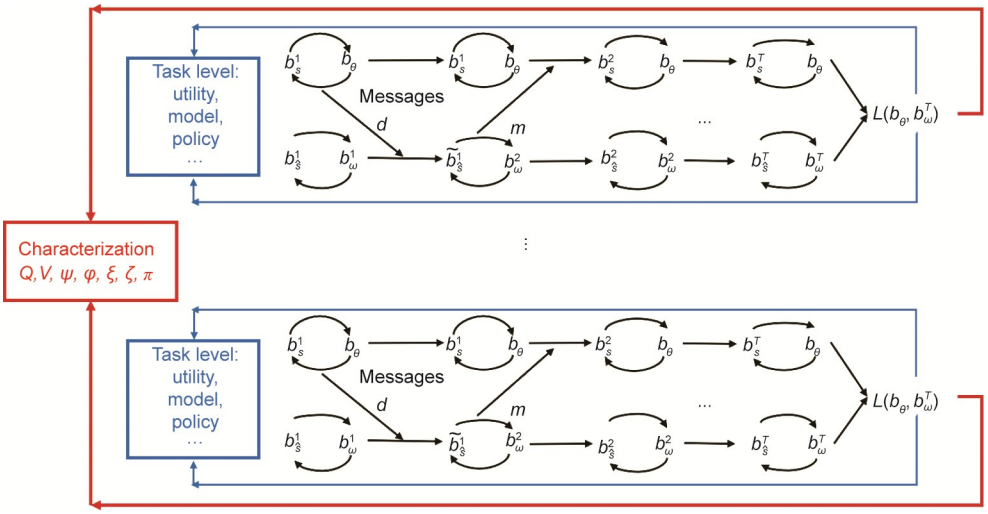

图17. 通学包括三个嵌套的循环：①黑色的反思循环（reflection loop），用于理解每一条信息，以实现共识；②蓝色的学习循环（learning loop）是为了实现共同的模型、效用、策略等；③红色的特征循环（characterization loop）是为了实现对每个智能体的特征的更好理解，如在3.3节中定义的价值和信念更新函数。当团队成员对其伙伴的特征描述的估计稳定下来时，通讯的规范（communication norm）就形成了。b_θ和b_ω是模型的信念；b_S、$b_{\hat{S}}$、$\tilde{b}_{\hat{S}}$是对学习状态的信念；m和d分别是来自学生和老师的信息；L表示损失函数，上标表示时间戳。所有的符号都遵循3.3节中的定义。

信念$b_S/b_{\hat{S}}$。

在这个层面上，理想的智能体应该能够捕捉到信息的字面意义和言外之意。到目前为止，大多数工作都假设老师和学生对所有信息的字面意义有共同的认识。换句话说，他们说的是同一种语言，或者至少，每个人都有一本这个语言的词典。若要弱化这一假设，可以拿走他们的字典，看他们是否能够从无到有地发展出一个有效的字典，作为信息的共同字面意思。这一层次的另一个挑战是理解言外之意，这往往需要反事实推理（counterfactual reasoning）。也就是说，为了理解说话者，听者不仅需要考虑收到的信息，还需要考虑没说的信息。当信息空间很大时，如汉语这样的自然语言，这种反事实推理往往很难精确求解，需要相应的近似方法。

5.2.1.2. 任务层次。任务层次的学习包括传递和理解一连串的信息，以获得老师和学生思维的协调。通学模式中的学习循环（learning loop）旨在找到这一层次的平衡点。也就是说，学生的效用、模型和策略要足够接近老师。为了评估收敛性，智能体需要直接或间接地测量他们思维之间的距离。直接测量依赖于各个思维成分的表示，如相对熵[Kullback-Leibler (KL) divergence]，适用于有闭式表达的信念。间接测量可以比较老师和学生之间的任务表现来，比如他们的运动轨迹之间的充分统计量（sufficient statistics）差异。我们在图17中把这个评价指标表示为$L(b_\theta, b^T_\omega)$。

由于教学通常需要一系列的信息，任务层次的学习通常需要做长远的顺序计划（sequential planning）。由于计划的复杂性与步数呈指数关系，而且信息空间通常很大或者甚至是是动态的，因此大多数老师都使用启发式教学法（heuristic）或只做短程计划。尽管如此，即使是这种不充分的计划，对于学生的建模也经常是局限于被动学习的学生（第一级范式）。第二级范式中的老师应该如何做计划仍然是一个方兴未艾的研究问题。它要求老师和学生形成一个团队的通讯规范（communication norm），这属于团队层次的范畴。在学习的任务层次还有另一个潜在的研究方向，那就是推广信息层次的语用学推理。也就是学生不仅可以从选定的教学信息序列中，还要可以从未选定的教学序列中学到知识。

5.2.1.3. 团队层次。最后，团队层次需要对老师和学生的特征（characterization）进行分析。要进行有效的通学，老师和学生的价值和信念更新函数必须要协调配合。在大多数已有的学习范式中，智能体的特征是由启发式方法先定的。在更一般和现实的环境中，老师和学生需要学习适当的特征来进行合作。通学中的特征循环（characterization loop）试图解决这个问题。就像人类团队需要大量的合作经验来获得默契，通学模式也需要多个学习任务来让智能体互相适应彼此的特征。如图17所示，多个学习任务的结果会塑造智能体的特征。4.1节中的指代游戏就遵循这个过程。在特征确定之后，老师和学生之间将建立通讯规范和学习协议。

本文中所提到的工作中，老师都是专门为一个或一个类型的学生设计的。一个特征循环也只为一对特定的师生发展学习协议。更一般的设定是让一个老师能够适应性地

教不同种类的学生，甚至是那些她在训练期间从未遇到过的学生。假设我们可以描述学生的特征，如智商、记忆力，那么一个因材施教的老师应该能够识别她的学生的性格，并相应地定制她的教学方法。具体来说，老师（学生）可以将她（他）的动态函数$\psi\varphi(\xi,\zeta,\pi)$参数化，并对参数的分布进行建模。这样的设置类似于任意组队（Ad-Hoc teaming [123]）和在具有不同动态函数的马尔可夫决策过程（Markov decision process, MDP）中进行多任务/元学习（multitask, meta-RL [124]），但在合作性教学的背景下的类似研究还没有广泛开展。

5.2.2. 停机的条件

了解了学习的三个层次之后，我们可以讨论学习的适当停止条件。在每个层次中，通学循环都会在不同的条件下终止。在信息层面，一个合理的终止标准是完全理解每条信息的字面意义和言外之意，即语用学含义。当智能体使用相同的信息空间时，理解字面意义通常很容易。语用学含义则取决于语境，如对话发生时的状态、通信历史等。由于语用学的理解需要心智理论，当递归推理收敛，或者进入下一层递归的认知负担超过了再发送一个信息来确认或澄清的成本时，反思循环就会停止。

在任务层次上，当六套思维收敛成一个，并由上帝的思维验证之后，学习就可以停止了。如果沿用图1中的记号，这就是说：

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

这是最严格的停机条件。在许多情况下，我们会定义其他停止条件。例如，让$D(X,Y)$表示两个思维X和Y之间的距离，比如相对熵、全变差距离（total variation distance）、动土距离（earth mover's distance）等。之后，我们有以下停止的条件。

• $\mathcal{D}(P_t,\hat{Q}_t)\leq\epsilon$，也就是说，老师认为学生已经知道她所知道的东西了，并停止教学。

• $\mathcal{D}\left(Q_{t-1},Q_t\right)\leq\epsilon$，即学生认为自己无法获得新知识。因此停止了学习。

• $\mathcal{D}(\hat{Q}_{t-1},\hat{Q}_t)\leq\epsilon$，老师认为学生不能再取得显著的进步，于是停止教学。

• $\mathcal{D}(C_t,\phi)\leq\epsilon$，老师和学生很难达成共识并终止了学习。

• $\mathcal{D}(Q_t,G)\leq\epsilon$，即学生在现实世界中取得了令人满意的表现，并停止学习。请注意，在某些情况下，G可能无法直接获得，因此可能需要某些代理函数。例如，让学生完成一些未见过的任务作为测试。

上述条件绝不全面，根据不同的需求和情况我们还可以提出更多的条件。此外，还可以定义比单一距离函数更复杂的条件，例如，比较最小化学生和老师之间分歧的收益与传输教学信息的成本。

在团队层次，老师和学生在了解了他们伙伴的特征，即动态函数、状态、模型和价值空间等之后，就可以停止学习了。这通常需要在多个教学任务上合作过之后才能完成。这个收敛的评估需要老师和学生在之前训练中没有见过的学习目标上进行教学。只有当他们能够在多个任务中进行有效的合作时，特征循环才可以结束并停止团队层次的学习。如果我们还希望有一个可以因材施教的老师，能够迅速适应各种学生。那么，停止的条件将是老师能成功地与多个学生进行通学，而且每个学生都需要进行多个学习目标的教学，类似于在随意组队问题中使用的评价方法[125]。

6. 结论

在本文中，我们提出了通学模式，它从通讯的角度来审视学习。我们回顾了现有的学习范式，如被动学习、主动学习和算法式教学，并研究了它们的局限性。新通学模式可以克服这些限制，并将先前的机器学习算法整合到一个统一的框架中。通过具体的例子，我们展示了高效的学习协议是如何从通学模式中从无到有产生的，并验证了该模式在复杂的人机交互任务中的实用性和必要性。此外，通学模式对机器学习基础理论有两个贡献：首先，它引入了新的学习表征，并提出了超越香农通信极限的学习协议；其次，它揭示了普遍的学习停机问题，指出了学习中的三个层次，并确定了可能的停止标准。总而言之，我们认为通学模式从老师和学生之间相互推理的角度为现有和未来的机器学习方法提供了统一的框架，并为未来开发更高级的学习范式打下了基础。

Acknowledgements

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-monthold 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] CausalityPearl J.. 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 semisupervised 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, VCdimension 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 semisupervised learning. In: Chapelle O, Schölkopf B, Zien A, editors. Semisupervised 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.