9

Teaching Robots New Tasks through Natural Interaction

Joyce Y. Chai, Maya Cakmak, and Candace L. Sidner

Abstract

This chapter focuses on the main challenges and research opportunities in enabling *natural interaction* to support interactive task learning. Interaction is an exchange of communicative actions between a teacher and a learner. Natural interaction is viewed as an interaction between a human and an agent that leverages ways in which humans naturally communicate and does not require prior expertise. The goal of communication is to achieve common ground and allow the learner to acquire new task knowledge. This chapter outlines the different types of knowledge that can be transferred between agents and discusses the perception, action, and coordination capabilities that enable teaching–learning interactions.

Introduction

Extending the framework introduced by Mitchell et al. (this volume), our focus in this chapter is on *natural interactions* between a human and an agent that enable interactive task learning (ITL). Reflecting most prior work on this topic, we focus on ITL scenarios where the teacher is a human and the learner is a physically embodied agent (e.g., robot) as opposed to a software agent.

Imagine an elderly couple, Katie and John Smith, who purchased a robot "Mia" as their personal assistant. Mia comes equipped with general knowledge of household chores and perceptual capabilities to recognize common household objects, such as those sold in grocery and hardware stores. Mia also has basic manipulation skills like grasping common objects or opening different types of containers. Despite these preexisting capabilities, Mia is unable to perform many tasks at Katie and John's house right out of the box. Not only does Mia need to be taught the unique tasks that the Smiths desire, it must also acquire new knowledge and capabilities to enable those tasks. The process of learning these tasks as well as task-relevant knowledge and capabilities

From "Interactive Task Learning: Humans, Robots, and Agents Acquiring New Tasks through Natural Interactions,"
edited by K. A. Gluck and J. E. Laird. Strüngmann Forum Reports, vol. 26,
J. R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978-0-262-03882-9.

happens through various forms of interaction with people, as in the following scenarios:

On the day of delivery, David, an employee from the company that manufactured Mia, arrives at the Smiths' with the new robot. David has an associate degree in robotic technology and has completed training on how to teach robots. The process starts with teaching Mia a map of the Smiths' house. David manually drives Mia to different rooms to construct the map and verbally provides information about each room as well as different points and regions in the room, such as where the main entrance is and the locations of appliances, trash bins, tools, and supplies. Next, David programs a set of basic skills tailored for the Smiths' house: how to open or close their cabinets, drawers, and appliances, for example, as well as how to operate various tools and appliances. He teaches Mia these skills by moving the robot's arm to demonstrate them. Then, under various scenarios, David tests the learned skills to ensure they are robust.

Once Mia is settled in the new house, the Smiths continue to teach Mia new knowledge and tasks. For example, they show where to put groceries or kitchen tools through pointing and verbally describing their locations with natural language: "The waffle maker goes in the bottom cabinet next to the stove." Katie also teaches Mia how to make their favorite dish from a family recipe. Using natural language and deictic gestures, Katie shows Mia different ingredients and demonstrates how and in which order to mix the ingredients. Mia sometimes has difficulty understanding Katie's instruction. For example, when Katie asks Mia to "grind the onion," Mia does not understand what "grind" means and subsequently asks for further instructions. Katie then provides detailed step-by-step instructions to show Mia how to perform the action "grind": "cut the onion in half, put the pieces into the blender, and press the top button." By following Katie's instruction and observing the change of the onion, Mia learns the meaning of the verb "grind" with respect to how the corresponding action changes the physical world. Mia can now transfer this understanding and perform related actions, such as "grind the carrot," assuming that Mia understands what a carrot is. Through this type of interaction, Mia continuously optimizes its task performance based on feedback from Katie, such as: "That looks slightly overcooked. Try reducing the baking time next time around."

For outdoor chores (e.g., a simple car maintenance task), John instructs Mia similarly to how he taught his son: he demonstrates how to (a) open the hood of the car, (b) check the engine oil, (c) check the radiator coolant and fill if needed, (d) check the windshield wiper fluid and fill if needed, and (e) replace the air filter if it is dirty. John and Mia both use language and deictic gestures to establish shared attention during the teaching–learning process. Once John explains and demonstrates how to fill radiator coolant, Mia can apply the learned skill to fill windshield wiper fluid. To teach the task, John uses conditional statements (e.g., "if the oil is below this line, then add coolant")

From "Interactive Task Learning: Humans, Robots, and Agents Acquiring New Tasks through Natural Interactions,"
edited by K. A. Gluck and J. E. Laird. Strüngmann Forum Reports, vol. 26,
J. R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978-0-262-03882-9.

and purposive descriptions (e.g., "you hold it because the funnel is too big," "put it so that the screw comes through the narrow part," or "place it right where the middle center opens into the screw so that the screw goes through the middle hole where it's open"). Mia extracts causal effect relations and converts them into schemas to support action planning and execution. The process also involves learning background knowledge mentioned in conditional statements, such as a too large funnel, the air filter being dirty, the time needed to hold an object in place, or the colors of objects through demonstrations or examples.

To understand Mia's capabilities and limitations, the Smiths can ask Mia different questions about its knowledge and its representation of the shared environment and tasks. These questions not only include "what" questions, but also "why" and "how" to assess Mia's reasoning and decision-making capabilities. Mia also proactively communicates with the Smiths about its internal representations of the world and the tasks, as well as the underlying reasoning that might take place to reach certain conclusions or decisions. Mia can even teach the Smiths' grandson how to cook their favorite dish and how to do car maintenance.

These scenarios illustrate different types of natural interaction that humans can use to teach robots new tasks or task-relevant knowledge and capabilities: by performing the task themselves, by verbally or kinesthetically guiding the robot, or through situated language instructions and gestures. This natural interaction between humans and agents instantiates the general framework of ITL, as shown in Figure 9.1. The human teacher has some *target task knowledge* in mind and intends to transfer this knowledge to the robot through various forms of interaction. Let S represent the set of states of the physical world relevant to the task and S_c represent the set of states of communication, such as the verbal utterances or focus of attention of the teacher at each step of the interaction. The robot learner perceives a *task-related world state* $s \in S$ through its sensors and constructs a *communicative state* based on its perception of the teacher's communicative actions. Let A represent the set of task-related actions (e.g., pick up an object) and A_c the set of communicative actions (e.g., asking for confirmation for its interpretation of a world state) available to the robot through its effectors. At each step of the interaction, the robot needs to decide what *task-related actions* $a \in A$ and/or *communicative actions* $a_c \in A_c$ it should take, given its current state and learning goals. The sequence of states and actions that a robot goes through during ITL constitutes its *interaction experience*. The robot needs to then extract *learning experience* from its interaction experience to obtain examples, specifications, and feedback to acquire new *task knowledge*.

Enabling ITL in robots through natural interactions requires a wide range of capabilities for perception, action, reasoning, learning, decision making, and communication. Here, we discuss the challenges and open questions associated with these capabilities. Specifically, we explore

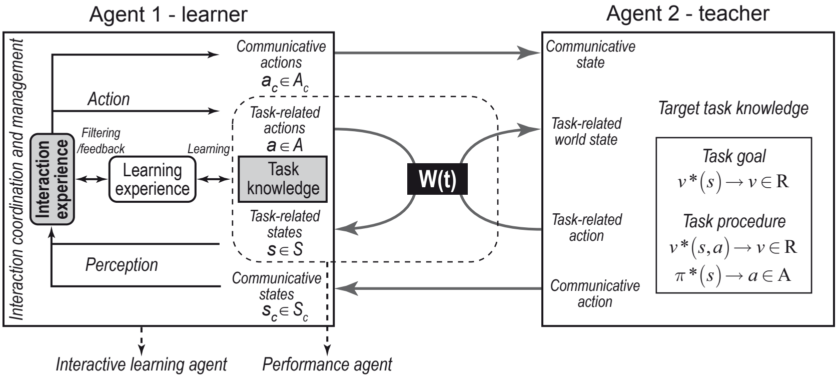
Agent 1 - learner
Agent 2 - teacher
Interaction coordination and management
Communicative actions $a_c \in A_c$
Action
Filtering /feedback
Learning
Task-related actions $a \in A$
Interaction experience
Learning experience
Task knowledge
Task-related states $s \in S$
Perception
Communicative states $s_c \in S_c$
W(t)
Communicative state
Task-related world state
Task-related action
Communicative action
Target task knowledge
Task goal
$v^*(s) \rightarrow v \in \mathrm{R}$
Task procedure
$v^*(s,a) \rightarrow v \in \mathrm{R}$
$\pi^*(s) \rightarrow a \in \mathrm{A}$
Interactive learning agent
Performance agent

Figure 9.1 Extended two agent and world model separating task-related states and actions from communicative state and actions. Task-related states (S) and actions (A) are the minimal set of states and actions that an agent needs to perform the target task successfully. Communicative states (S_c) and actions (A_c) are what an agent needs to communicate to extract useful data and provide feedback for learning.

- forms of human teaching and the different kinds of knowledge that can be taught or learned through interaction,
- capabilities to perceive and infer task-related state and communicative state through sensors, including visual scene understanding, language understanding, and grounding language to visual perception (e.g., the environment, perception of human gestures, and perception of human actions),
- capabilities to act in the environment through effectors, including acting to manipulate the environment and communicating to the human during interaction, and
- capabilities to manage and coordinate interaction and establish common ground.

Types of Task Knowledge and Forms of Interaction

Humans can learn new tasks from other humans through various means: watching each other perform the task, doing the task themselves accompanied by instructions and guidance, or conversing and imagining the task without performing any actions (e.g., acquiring a new recipe). Similarly, as illustrated in our example scenario, robots can learn from humans in analogous ways.

As shown in Figure 9.2, in ITL the robot needs to extract learning experience from interaction experience through interaction. The learning experience can involve examples of goal states, examples of action sequences that lead to a goal, or evaluations of action sequences generated by the robot. These

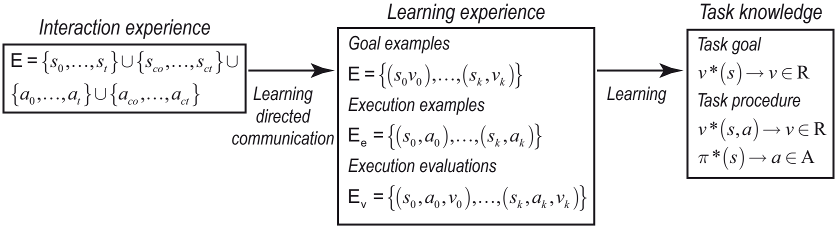
Interaction experience
$E = \{s_0,\ldots,s_t\} \cup \{s_{co},\ldots,s_{ct}\} \cup \{a_0,\ldots,a_t\} \cup \{a_{co},\ldots,a_{ct}\}$
Learning directed communication
Learning experience
Goal examples
$E = \{(s_0 v_0),\ldots,(s_k,v_k)\}$
Execution examples
$E_e = \{(s_0,a_0),\ldots,(s_k,a_k)\}$
Execution evaluations
$E_v = \{(s_0,a_0,v_0),\ldots,(s_k,a_k,v_k)\}$
Learning
Task knowledge
Task goal
$v^*(s) \rightarrow v \in \mathrm{R}$
Task procedure
$v^*(s,a) \rightarrow v \in \mathrm{R}$
$\pi^*(s) \rightarrow a \in \mathrm{A}$

Figure 9.2 Interactive task learning is the process of converting the learning agent's experience into task knowledge. Different types of knowledge are learned from different types of example data.

different learning experiences can be expressed in terms of the physical world state (s_i), task-related actions (a_i), and values assigned to them (v_i). The goal of task learning is to extract different types of task knowledge such as task goal (e.g., $v(s) \rightarrow v$) and task procedure (e.g., a policy to perform the task $\pi(s) \rightarrow a$) from these experiences. Different learning algorithms require specific types of experience data (e.g., direct policy learning requires sequences of state-action pairs). The role of the communicative actions is to extract this data from the unstructured stream of data that the agent experiences. For instance, communicative actions by the teacher might indicate the start and end of a demonstration to help the learning process, even though the communicative states and actions are excluded from the learning data. As we discuss next, the way in which task knowledge is transferred and the role of communicative actions in that process largely depends on the type of task knowledge.

Task Knowledge Types

The main goal of task learning is to acquire *task knowledge*, which defines what a task is and provides sufficient information to permit the robot to perform the task on its own. There are different types of task-related knowledge and capabilities (described below) that can be acquired during interaction. As discussed by Laird et al. (this volume), task-related knowledge often includes goals, actions, and constraints which define the problem space as well as procedural/policy knowledge needed to perform the task. Here, we focus on two types of task knowledge and their representations: task procedures and task outcomes.

Task procedure information captures what the agent needs to do to complete the task, as shown in Figure 9.2. Most existing agent frameworks represent procedural information as a policy, which is a function that maps the perceived state to an action: $\pi(s) \rightarrow a$. Such functions can be represented by many different types of classifiers or regressors and be learned from examples. Process information can also be captured in more explicit forms such as plans, programs, finite-state machines, or hierarchical task networks. Although these different representations do not necesarily provide a full mapping of states to actions, they still capture procedural knowledge by specifying a sequence, a partial ordering, a schedule, or a hierarchical organization of actions in the context of a task. For example, Pardowitz et al. (2007) introduced task precedence graphs that capture ordering constraints between actions involved in a task. Similarly, Ekvall and Kragic (2008) represent tasks with a set of ordering constraints between pairs of actions. Alexandrova et al. (2015) use a flow diagram to represent tasks with actions that have pre- and postconditions that can cause branching in the program. Huang and Cakmak (2017) use the general-purpose visual programming language, Blockly, to represent various branching and looping tasks.

Task outcome information relates to the goals or desired outcomes of a task, independent of the process followed to achieve them. This is different from the actual outcomes when performing a task (which can be expected or unintended). Task goals are often captured by the reward or value functions associated with states and actions, assuming the agent is maximizing reward or value. In practice, task goals might be easier to express in terms of world states in which the task is considered complete; for instance, a conjunction of state variables that need to be true or other arbitrary functions that evaluate a given state in terms of whether the goal is achieved. A value can then be associated with each state based on how close they are to a goal state. The task "tidy up the living room," for example, could be specified with the list of items in the room and their desired locations, without any information on how to get them there. Such a representation was used by Chao et al. (2011) to represent simple object reconfiguration tasks. The ability to carry out tasks based solely on specified goals often requires the robot to have planning capabilities.

Some task representations involve combinations of process and outcome information. For example, a recipe for a particular dish specifies not only a sequence of actions but also mentions what to expect at the end of the process or when a task is considered complete.

Forms of Interaction in Transferring Task Knowledge

There are many forms of interaction that enable transfer of task knowledge. Our focus here is on two key types of information transferred in those interactions: *demonstrations* of the task and direct *specifications* of task constraints or properties.

In learning task processes from *task demonstrations*, multiple demonstrations provide alternative ways of achieving the same task (e.g., Argall et al. 2009). Different task representations capture this information in different ways. For example, a partially ordered plan captures alternative orderings of low-level actions. Hence different demonstrations of the task might involve a different ordering of actions. Similarly, a program with conditionals and loops

From "Interactive Task Learning: Humans, Robots, and Agents Acquiring New Tasks through Natural Interactions,"
edited by K. A. Gluck and J. E. Laird. Strüngmann Forum Reports, vol. 26,
J. R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978-0-262-03882-9.

captures alternative ways of performing a task, depending on a perceivable "condition" or different numbers of repetitions contingent on user-specified or environmental parameters. Different demonstrations of such tasks will involve alternative traces of the program. Task outcomes can also be taught by demonstration. Multiple examples provide variations of the states in which the task is considered successfully completed. One of the key computational challenges is to identify parts of the state that are relevant/irrelevant for the task. Thus it is important for demonstrations provided by the teacher to involve such variations.

Tasks can be demonstrated through different forms of interaction (e.g., by the teacher performing the task, or provided directly to the robot, with guidance from the human teacher). In tasks performed by humans, one of the most intuitive ways to demonstrate a task is for a human to perform it herself. For the robot to learn from this type of demonstration, the robot must be able to perceive the human's actions and/or the effects of human action on the environment. Perception of human actions can be facilitated through external sensors or wearable sensors on the human. Once a robot perceives human actions, they need to be mapped to corresponding robot actions. This is referred to as the retargeting problem. In some cases, perception of actual actions is not necessary, as long as the robot can detect the state changes that result from the task demonstration and learn the task based on that information (Baisero et al. 2015; Mollard et al. 2015).

For tasks performed by a robot with guidance from a human, the human teacher demonstrates a task to the robot by guiding it through the task. This mode of teaching bypasses the retargeting problem but requires the teacher to have a good understanding of the robot's action capabilities. The guidance to the robot can be provided in various ways, from kinesthetic movements to verbal instructions:

- *Kinesthetic guidance* involves physically holding the robot and moving its manipulators to perform the task (e.g., Akgun et al. 2012; Phillips et al. 2016).
- *Natural language guidance* involves instructing the robot on what to do to perform the task. Mohan et al. (2012) and She et al. (2014), for instance, use step-by-step language instructions to teach new tasks to a robot.
- *Multimodal language guidance* uses multimodal instructions (speech and gestures) to guide a robot through the task.
- *Gestures* often serve to reference parts of the environment.
- *Joystick-based guidance* involves driving the robot and triggering prespecified actions with the help of a special device to perform the task.
- *Guidance based on graphical user interfaces* (GUIs) employs a graphical interface to control the robot and trigger prespecified actions to perform the task.

In *task specifications,* alternative ways of achieving a task are directly specified by the teacher in a format compatible with the robot's task representation. For example, for a partially ordered plan representation, the teacher might verbally state:

> First *bring all of the ingredients and tools* to the kitchen counter (in any order). Second, pour *all of the dry ingredients* into the mixing bowl (in any order).

While teaching by demonstration inevitably involves a particular ordering of the actions and hence requires multiple demonstrations to capture order invariances, direct specification provides an efficient way to provide the same information. Similarly, if the representation is a program, the user can directly specify loops or conditionals by literally writing a program or verbally specifying those with instructions:

> Insert a toothpick into the center of the cake. If it comes out clean, take out the cake; otherwise continue to bake. Alternatively, for each cup on the muffin pan, pour until three-quarters full.

Similarly, direct specification of task goals involves the teacher directly indicating parts of the world state that are relevant or irrelevant to the robot's task, rather than trying to exemplify variations of positive and negative goal states. For example, the teacher may verbally describe the desired goal state when teaching a robot to set a table:

> The red bowl should be on top of the green plate, and the napkin should be placed to the right of the plate.

Task specifications can be provided through natural language or GUIs:

- *Natural language specifications* involve the use of language to specify directly certain properties or constraints about the task representation. For example, Cantrell et al. (2012) use natural language to specify precondition and effects of action schemas for task planning.
- *GUIs* can be used to specify properties or constraints about a task being taught to a robot.

Humans often combine these two means of communicating task knowledge (demonstrations and specifications). For example, a teacher might demonstrate the physical act of adding different ingredients to the mix in a particular order as part of teaching a recipe, while verbally specifying partial ordering constraints by saying: "Add all dry ingredients in any order." Similarly, a person might set up the table themselves to show an example of how they want the table to be set, but then specify invariance constraints by saying: "The salt and pepper can be anywhere in the center area of the table."

Regardless of the specific form of interaction, during learning, symbolic representations of human inputs (e.g., GUI, natural language) need to be tightly grounded to the robot's internal representations of perception and action.

Task-Relevant Background Knowledge and Capabilities

When we speak about a robot learning new tasks, we often assume that the robot has the necessary background knowledge and capabilities. The ability to perform new tasks, however, might equally be due to the acquisition of other knowledge or capabilities, not solely due to newly acquired task knowledge. Hence, the ability to acquire these different kinds of background knowledge and capabilities through interactions is also highly relevant for task learning. For instance, the capabilities of a robot that already knows the task of sorting objects, based on different properties, can be expanded through the acquisition of new perceptual capabilities (e.g., the ability to detect new object properties) or new action capabilities (e.g., the ability to manipulate new types of objects). Below we list four types of knowledge and capabilities relevant for task learning:

1. *Perception capabilities* refer to the ability to perceive the task-relevant environment and interpret human language, including
   - state and actions of humans,
   - state, properties, and affordances of objects,
   - scene composition (surfaces, objects, humans, and their relationships),
   - changes of the state that occurred to the environment, and
   - state of communication such as communicative intent and focus of attention.
2. *Action capabilities* refer to lower-level policies that control a robot's actuators to carry out tasks and/or communicate with humans. These include capabilities that allow robots to
   - navigate the environment,
   - manipulate objects in the environment, and
   - communicate with humans in the environment.
3. *Linguistic knowledge* concerns the meanings of words and phrases. For physical robots, which need to sense from and act upon the physical world, as opposed to the symbolic world, this knowledge cannot be purely symbolic as in a dictionary or thesaurus. Word semantics need to be grounded to the robot's sensorimotor skills.
4. *World knowledge* captures any other task-relevant knowledge about the world and how the world works:
   - *Facts* about the world and the robot's task environment: "my owner's name is Katie Smith" or "I was built in 2017."
   - *Commonsense knowledge* that allows a robot to interpret human language and reason about its perception and action goals (Al-Moadhen et al. 2013; Tenorth and Beetz 2009): to "boil the water," water must first be placed in the boiling pot.

- *Action knowledge* that captures the existing knowledge about subtasks and subgoals previously acquired or learned. Formal action models capture preconditions and effects of actions (Fox and Long 2003). Preconditions specify world states in which the action is applicable; effects specify the expected changes to the world state.
- *Domain knowledge* that corresponds to information specific to a particular task environment or user that a robot needs to perform its task. For example, a robot that performs object deliveries to hotel rooms needs to have a map specific to the hotel within which it is deployed, with room numbers annotated on the map.

Some of these types of knowledge and capabilities can be programmed into a robot. They can also be acquired through interactions with humans, although the means of acquisition is less clear than that for task knowledge.

Forms of Interaction for Learning Task-Relevant Knowledge or Capabilities

The types of interactions that support acquiring task-relevant knowledge and capabilities are similar to those involved in learning the task itself. As shown in Table 9.1, the forms of interaction often depend on the kind of knowledge or capabilities to be learned. For example, to help train the robot's visual perception capabilites, the teacher may use language descriptions and also show target objects from different angles. To acquire the navigation map, teleoperation (e.g., through joystick guidance) can be employed as well as language

Table 9.1 Example forms of interaction for different types of knowledge

Knowledge	Example Forms of Interaction
Perception capabilities	• Natural language and deictic gestures to teach labels of objects and indicate their relations • Natural language to specify object affordances
Action capabilities	• Kinesthetic demonstration to teach low-level control policies to generate arm trajectories or navigation strategies
Linguistic knowledge	• Natural language combined with deictic gestures to teach nouns and adjectives • Natural language combined with action demonstration to teach action verbs
World knowledge	• Natural language to specify order constraints among sub-actions • Natural language to specify causality (i.e., precondition and effect) of an action • Demonstrations performed by the human to show how basic actions/verbs change the state of the world • Joystick guidance to build a map of the robot's environment for navigation

From "Interactive Task Learning: Humans, Robots, and Agents Acquiring New Tasks through Natural Interactions,"
edited by K. A. Gluck and J. E. Laird. Strüngmann Forum Reports, vol. 26,
J. R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978-0-262-03882-9.

descriptions. Acquisition of low-level action knowledge (e.g., lower-level policies to generate trajectories) may benefit from kinesthetic demonstration whereas higher-level task knowledge (e.g., partial orderings) may best benefit from language instructions. Linguistic knowledge certainly involves the use of language, which is often combined with deictic gestures or action demonstrations because the semantics of words need to be grounded to visual perception and the change of state in the physical world.

Open Questions in Enabling Effective Task Learning Interaction

Teaching Presupposed Task-Relevant Knowledge

While previous work has investigated the acquisition of many types of task-related knowledge and capabilities, the acquisition of commonsense world knowledge in task learning has largely gone unexplored. In human–human interactions, knowledge about the world and the domain is often presupposed. The speaker and the listener believe they share the same kind of world knowledge, so it does not need to be explicitly stated. However, in human–robot interactions, huge discrepancies in world knowledge can exist between humans and robots. Often, the robot does not have sufficient background knowledge to learn a new task. Thus human teachers need to be able to assess what kind of background knowledge the robot has and how to teach the robot background knowledge pertinent to the task at hand. For example, the result states of basic action verbs are not usually specified, and humans naturally take them for granted. Existing lexical resources (such as Verbnet, FrameNet) and preexisting knowledge bases (e.g., Google's Knowledge Graph, Freebase) do not offer the level of detail required for the robot to understand the very basic principles about the conditions for their actions (e.g., "put A on B" requires A generally smaller and lighter than B) and how their actions may change the world (e.g., slicing a cucumber may lead to the change of the shape, size, and pieces of the cucumber).

Thus, it is important to understand *what a human must teach a robot about the domain of a task.* Some background knowledge (e.g., time as duration, units of time, and time relations) may be best taught or acquired once for many domains, but much human knowledge is domain specific. Learning domain-specific knowledge leads to a whole new set of research questions:

- How does the human know what knowledge the robot (e.g., subactions) has so that it can be used to teach new tasks?
- During task learning, what signals indicate the lack of background knowledge and therefore human teaching is required?
- How can existing resources be leveraged to acquire the correct level of background knowledge during teaching?

From "Interactive Task Learning: Humans, Robots, and Agents Acquiring New Tasks through Natural Interactions,"
edited by K. A. Gluck and J. E. Laird. Strüngmann Forum Reports, vol. 26,
J. R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978-0-262-03882-9.

- What level of granularity should background knowledge be taught by a human?
- How should background knowledge be represented and used for effective reasoning and inference?

Combining Different Forms of Interaction for Task Learning/Teaching

Most previous work on task learning has focused on a single form of interaction for teaching. Except for a limited few (Kirk et al. 2016; Mohseni-Kabir et al. 2018; Niekum et al. 2015; Rybski et al. 2007), techniques that combine language, dialogue, and action demonstration to teach complex tasks are in critical need. As discussed above, different forms benefit different types of knowledge. In addition, as the situation changes (e.g., the lighting situation changes from being good to poor), the form of interaction may need to adapt (e.g., switch from visual demonstration to language instruction). Thus we need to know *how to seamlessly combine and adapt different forms of teaching to enable the most effective teaching*. Is combining and adapting a problem for human teachers or a problem for robot learners? The answer is both.

Teaching Humans How to Teach Robots

After working with a robot, an experienced human teacher (in our scenario involving Mia, this would be David, the employee from the robotic manufacturer) should be able to discern which form of interaction is necessary to teach a specific kind of knowledge to meet specific circumstances. Experienced human teachers should know when to provide a particular kind of feedback (i.e., reward or punishment) so that the robot can learn from such feedback and adjust its behaviors to maximize future rewards. Experienced human teachers may also apply scaffolding, intentionally vary the situation, and design different experiences for the robot to learn the task and aspects associated with the task.

Thus, similar to the setting in human skill learning, human teachers' behaviors and experience have a massive influence on the success of robot task learning. *How, then, should we train a new generation of human partners/teachers, so that robots can be effectively taught through their collaborations*?

Enabling Robots to Engage Proactively in Learning

We cannot expect that every human partner will be capable of identifying and employing the most effective means to teach the appropriate kind of knowledge. Thus a robot needs to be able to share the burden of selecting effective strategies. A crucial issue, as yet unstudied, is: *How can a robot be made to be aware of its own learning situation—one in which it is capable of communicating to the human its limitations and proactively requesting the right kind of teaching from the human?*

From "Interactive Task Learning: Humans, Robots, and Agents Acquiring New Tasks through Natural Interactions,"
edited by K. A. Gluck and J. E. Laird. Strüngmann Forum Reports, vol. 26,
J. R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978-0-262-03882-9.

Capabilities to Perceive the Environment and Human Inputs

The ability to perceive the environment and human inputs as well as to infer current task-related states and communicative states is fundamental to ITL. A robot must be able to recognize task-relevant objects in the environment, the change of the environment caused by an action, task demonstration from humans, as well as verbal and nonverbal human communicative behaviors. It must also be able to infer human intent, interpret instructed actions and their involved objects, and derive task structures by grounding language to perception.

Visual Perception

Performing or learning tasks inevitably requires an understanding of objects and environments integral to the tasks. This includes objects, their properties, fluents (i.e., attributes which can potentially change), and relations, as well as an understanding of external actions and how they may have changed the perceived state of the physical world. As humans can perform actions to teach robots and apply nonverbal modalities (such as deictic gestures, iconic gestures, and gaze directions) to facilitate communication, the robot should also have the capability to recognize the state and actions of its human partners.

Acquiring perceptual capability has been the main research goal for the computer vision community. Most of the learning algorithms for perception are trained offline and rely on large training data for object recognition, activity recognition, and so forth. Recent years have seen significant progress on recognition of common objects from static scenes (e.g., images) (Grauman and Leibe 2011). However, in a dynamic scene, such as would be encountered in task learning, object tracking and human action recognition still face many challenges (for reviews, see Aggarwal and Ryoo 2011; Sargano et al. 2017). In addition, during task learning, it is likely that neither relevant computer vision models nor sufficient data are available. Thus, it is critical for the robot to continuously acquire new models for object recognition through interaction with its human teacher. The teacher can use language to provide the name, the object type, and related properties to a perceived object in the environment, and the robot needs to learn a generalized model efficiently (e.g., for object recognition) that can be applied in new situations. Key research questions include:

- How can a robot learn reliable models based on a small number of examples with limited human supervision during interaction?
- How can it transfer and adapt models learned from previous experience to a new situation (e.g., transfer learning), perhaps with limited human intervention?

From "Interactive Task Learning: Humans, Robots, and Agents Acquiring New Tasks through Natural Interactions,"
edited by K. A. Gluck and J. E. Laird. Strüngmann Forum Reports, vol. 26,
J. R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978-0-262-03882-9.

Language Understanding

Language plays an important role in ITL. From a human's linguistic utterance, the robot needs first to understand the underlying intent of the teacher (e.g., whether it is to teach the robot a new step or to correct the robot's current understanding of a learned step or action). When a referring expression is involved, the agent needs to understand what entities, from the interaction discourse or the shared environment, are being referenced. When the utterance describes some task steps, the agent needs to understand what actions are specified and what participants are involved (e.g., agent, patient, instrument, source, destination). The robot also needs to be able to extract any information from the utterance that specifies preconditions, effects, and constraints (e.g., temporal orders) associated with actions and tasks. To help achieve the abovementioned abilities, recent advances in natural language processing—particularly in syntactic parsing, semantic processing, and discourse processing—can be applied (Jurafsky and Martin 2008). In the event that the robot cannot successfully understand human utterances, dialogue can be applied to clarify human intent and disambiguate different interpretations of linguistic expressions.

In situated interaction, language communication is often accompanied by other nonverbal modalities, such as gesture. Deictic gestures (e.g., pointing to objects in the environment) and iconic gestures (e.g., waving hello or indicating an action or a particular type of object) are vital to an understanding of the teacher's intent. Pointing gestures are essential to task instruction because the array of objects in a task (which may be difficult to describe verbally) lead to the need to point at them rather than rely solely on language descriptions. Matuszek et al. (2014), for example, combine language and gesture to interpret directives in human–robot interaction.

Speech communication is perhaps one of the most natural means of interaction in task learning. Speech recognition has made significant progress over the last decade. More recently, advances in deep neural networks have made it possible for machines to achieve recognition performance on par with human performance. At the time of writing of this article, Google reported a 4.9% word error rate in recognition while human performance is estimated to be around 4% word error rate (Saon et al. 2016). Although encouraging, these results were often obtained based on offline benchmark data. Thus, it is not clear whether the same performance can be attained in a real-time, interactive, and unconstrained environment. *How can recent advances in speech recognition be successfully applied to real-time interactive systems for task learning*?

Unlike traditional natural language processing, linguistic knowledge must go beyond pure symbolic representations—as in a dictionary or thesaurus—to enable communication with physical robots. The meanings of words need to be grounded to the robot's internal representations that are connected with sensors and effectors. Concrete nouns, for instance, need to be grounded to the types of objects or object attributes perceived from the environment (e.g., color words grounded to color histograms). Adjectives are often grounded to the perceived attributes (e.g., the size of the bounding boxes, the weights of an object) and fluents (e.g., door open or closed, box open or closed). Verbs need to be grounded to the underlying action representations, which can be accessed by the robot's control system to plan and execute the corresponding actions. On one hand, existing knowledge of grounded word semantics will be applied to ground language to perception and action (discussed in the next section). On the other, as new words are often encountered during interaction, they should be acquired continuously through situated interaction (Mohan et al. 2012). When a situation changes (e.g., a change in the environment), the learned word representation may not fit the new situation (e.g., a lighting change in the environment may affect grounded word models for color words). Thus, it is important that word models are adaptable to new situations (Liu and Chai 2015; Thomason et al. 2015).

Grounding Language to Perception

The capability to ground human language to the perceived physical environment is particularly important for task learning. Suppose a human teaches the robot how to boil water by demonstrating to the robot how to achieve this task through step-by-step instructions: "pick up the pot, fill the pot with water, boil the water…" To learn how to perform this task, the robot must first understand what perceived objects are involved in each step of instruction by grounding the arguments of action verbs, such as the noun phrase *the pot*, to the perceived objects in the environment.

This task of grounding language to perception of the environment has received an increasing amount of attention (Krishnamurthy and Kollar 2013; Matuszek et al. 2014; Mooney 2008; Tellex et al. 2011, 2014; Yang et al. 2016; Yu and Siskind 2013). Most previous approaches first process language and vision separately, and then integrate the partial results together. In a dynamic scene with ongoing activities, computer vision algorithms still have difficulty reliably recognizing and tracking objects and actions; this leads to a bottleneck in grounding language to vision. Recent deep learning approaches directly fuse raw features from language and vision and have achieved state-of-the-art empirical results on applications such as caption generation from images/videos and visual question answering. These approaches, however, require a large amount of training data. *To integrate language and vision in the context of ITL, what would be the optimal architecture?*

Another line of recent work has explored causality modeling for action verbs (Gao et al. 2016). Here the idea is that knowledge of how concrete action verbs (e.g., cut, slice, pick up, etc.) might alter the world can drive visual detection. For example, from the directive "*slice the cucumber*," knowledge about expected changes to the cucumber will provide high-level guidance to look specifically for grounded objects with relevant features (or the change of features) in the visual scene. Recent work has also explored commonsense physical knowledge about objects that are implied by action verbs (Forbes and Choi 2017). For example, “*he threw the ball*” implies that “*he*” is bigger and heavier than “*the ball*.” This kind of implicit knowledge can potentially provide additional cues to ground language to perception.

Capabilities to Act and Communicate

Enabling a robot to learn new tasks requires capabilities to carry out task-related actions as well as actions that facilitate communication. These capabilities span a wide range, from navigation and manipulation to communication.

Task-Related Actions and Grounding Language to Action Representation

A robot’s action capabilities can be based on manually designed and tuned controllers as well as policies learned from human demonstrations or through reinforcement learning. In some robotic applications, it is essential for the robot already to possess all of the action capabilities needed to complete a task. For example, previous work in the robotics community aimed to translate natural language instructions to robotic operations (Kress-Gazit et al. 2007; Spangenberg and Henrich 2015), but they were not designed for learning new actions or tasks. In other cases, tasks and actions can be learned simultaneously. For example, Mohan and Laird (2014) developed a system where a robot can learn a hierarchical representation of a new task based on linguistic interaction with the human. Similarly, Liu et al. (2016) applied grammar induction to learn a hierarchical and/or graph representation for a new task from a human’s language instructions and visual demonstrations.

To support action learning from language instructions, recent work has begun to explore the connection between semantics of concrete action verbs and action planning (Misra et al. 2016; She et al. 2014) and explicitly represented grounded verb semantics as desired goal states of the physical world as a result of the corresponding actions. Such representations are learned based on example actions demonstrated by the human. For example, a human may teach the robot how to “boil water” by issuing step-by-step language instructions which the robot knows how to perform: “move to the kettle, grasp the kettle, move to the stove…” By following these steps, the robot will experience the change of the physical world. By capturing the differences between the goal state and the initial state, the robot is able to acquire the semantics of the verb frame “boil (water).” Once acquired, these grounded representations will allow the robot to interpret verbs/commands issued by humans in new situations and apply planning to execute actions. One limitation of previous work is that the algorithms were mainly developed based on simulations (e.g., simulated Baxter robots). Except for a few (e.g., She and Chai 2017), uncertainties from the environment were largely ignored. However, the world is full of uncertainties at various levels: from motion planning to perception and language grounding. To extend task learning from language instructions to the physical world, it is paramount to address *how to integrate uncertainties at multiple levels together, so that new actions associated with concrete action verbs can be learned*.

Verbal and Nonverbal Communicative Action

Separate from its task-related actions, a robot will need to perform communicative actions to facilitate its learning/teaching interactions. In situated interaction, both verbal and nonverbal modalities are available for the robot to communicate to its human partner. Some example communication abilities include:

- generating speech and deictic gestures to confirm understanding of instructions or refer to objects in the environment (Fang et al. 2015);
- generating gaze direction, communicative head gestures (e.g., nodding and shaking head), or facial expressions (confused or confident face) to respond to human input at different points in the interaction (Holroyd et al. 2011); or
- displaying visualizations of learned concepts to enable humans to inspect them.

In particular, the embodiment of a physical robot can take advantage of nonverbal modalities (e.g., gaze and gesture) for efficient communication. The robotics community has learned from psychologists that gazing at others and at objects in the environment are quintessential human behaviors. Gaze that is used to convey information to a collaborator is referred to as *social gaze*. Gaze at a collaborator functions to gather attention from the other, to indicate social presence, and to indicate attention to the individual (e.g., turn-taking via gaze aversion). Gaze at objects serves to indicate what one is paying attention to, is about to point at, what one intends to do next, or to indicate that what another has focused on should now be the object of mutual gaze. Collaborators use gaze information to assess how well their partners comprehend their collaborations as well as to assess the collaborators' level of continued engagement (Rich et al. 2010). Every one of these abilities is valuable in task learning, as they enable the assessment of how the learning is progressing, whether the learner is looking in the right direction, and what the teacher intends for the learner to do. Gestures also have similar effects in coordinating interaction, establishing shared attention, and providing feedback. Proxemics, which models the stance of individuals to others and how they approach one another, can be significant in tasks because where the learner stands in performing a task may be crucial. *How to generate verbal and nonverbal communicative behaviors effectively to facilitate task learning remains an important focus for future research.*

Capabilities to Manage and Coordinate Interaction

Managing interactions between humans and robots is critical to support task learning/teaching. At any point in the interaction, robots need to decide what to do next based on interaction history, current situation, and learning goals. These decisions can be made by following simple decision rules that are manually crafted or interaction policies that are learned from experience.

Interaction Management and Active Learning

Decades of work on dialogue modeling are relevant for ITL. Different approaches have been developed, for example, driven by intention and collaboration (e.g., Grosz and Sidner 1986; Rich and Sidner 1998), based on information states (Larsson and Traum 2000) or interaction policies learned from reinforcement learning (Kaelbling et al. 1996; Young et al. 2013). Despite recent progress, dialogue modeling remains a significant challenge. Dialogue models need to be able to accommodate interruption, turn-taking, and other dialogue behaviors, which neither the intention-based nor information state approach have successfully addressed, but are essential in task instruction.

Specifically, to learn new tasks, active learning has been shown to be an important component that contributes to effective interaction management. Most work on task learning assumes a learner that passively receives information from the teacher. However, humans are often suboptimal in their teaching when the learner is passive. One line of work explores active task learning whereby the learner actively requests specific information that it evaluates as most useful. Active questioning enables much more efficient learning. For example, Chao et al. (2010) and Cakmak et al. (2010) demonstrated that an active learner (both human and robot) which requests labels (positive/negative) for specific instances of a task goal outperforms a passive learner taught by examples selected by naïve human teachers. In particular, Cakmak and Thomaz (2012) identified three types of queries that can be used by a human/robot student as part of active task learning:

1. Demonstration queries asking for a full or partial demonstration of the task
2. Label queries asking whether an execution is correct
3. Feature queries asking about the relevance or invariance of specific aspects of the task

Recent work by She and Chai (2017) extended this question–answer style of interaction and applied reinforcement learning to acquire an interaction policy that allows the robot to handle noisy environment and learn new verbs and corresponding actions. To improve ITL, we need to know how to engage in a full range of interaction that can incorporate active learning with other

From "Interactive Task Learning: Humans, Robots, and Agents Acquiring New Tasks through Natural Interactions,"
edited by K. A. Gluck and J. E. Laird. Strüngmann Forum Reports, vol. 26,
J. R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978-0-262-03882-9.

communicative goals (e.g., clarification and disambiguation) to acquire more reliable models of skills.

Extra-Collaborative Effort and Transparency

In human–human task learning, human teacher–learner partners often share similar perceptual capabilities as well as basic commonsense knowledge to support their collaboration.

In human–robot task learning, however, there are huge discrepancies in background knowledge between humans and robots. The robot, for instance, often does not have sufficient background knowledge to learn a new task. Furthermore, although they may be co-present in a shared environment, humans and robots have mismatched capabilities in reasoning, perception, and action: their representations of the shared environment and joint tasks can be significantly misaligned. A significant challenge involves the lack of common ground and discrepancies in the human's mental model of what a robot knows and is capable of doing. Previous work (Chai et al. 2016) has shown that to bridge the gap and strive for a common ground of shared representations between humans and robots, extra effort is needed. This extra-collaborative effort in interaction not only has implications in algorithms for language grounding, but also affects interaction management.

Transparency plays an important role in achieving common ground and promoting accurate mental models during interaction. For example, Thomaz and Breazeal (2006) demonstrated that natural transparency mechanisms, like gaze, can steer the human's behavior while demonstrating a task. Pejsa et al. (2014) used facial expressions to provide transparency about dialogue uncertainties. Alexandrova et al. (2015) employed interactive visualizations of learned actions to enable teachers to verify tasks that are learned from a single demonstration and correct any mistakes they detect. Guha (2016) used pointing to communicate the robot's understanding of a referenced object, and Whitney et al. (2016) used heat map visualizations and facial expressions to communicate uncertainty about its inference. Recent work by Hayes and Shah (2017) allows a robot to automatically generate verbal description of its learned policy (i.e., which actions it takes in which contexts).

To enable common ground for effective task learning, there are many research questions to pursue:

- How can an agent make its internal representations (e.g., causal-effect relations) transparent to the human?
- How can an agent explain its autonomy or decision so that the human can better understand the agent's capabilities and limitations?
- What are the mechanisms to manage interaction so that it can encourage a human's collaborative behaviors and simultaneously create more collaborative behaviors from the robot?

Conclusions

To fully support teaching robots new tasks through interaction, many challenges and open questions remain as discussed above. While the scenarios in the introduction focused on in-home settings, teaching robots new tasks is applicable in many situations, especially ones with highly structured environments. Already robots are being trained by people in ad hoc ways to work in manufacturing assembly lines (e.g., Guizzo and Ackerman 2018). Robots working in warehouses are largely programmed by hand, but it is not difficult to envision the need for them to be taught tasks by human coworkers. The same applies to robots in the service industry (e.g., hotel helpers).

One key challenge in task learning, which we did not discuss, is evaluation—a critical and difficult issue in interactive systems because many confounding factors are involved. In the context of ITL, the following questions arise:

- How do we know that the task has been learned?
- What additional metrics are needed to evaluate the success of task acquisition beyond traditional metrics for evaluating interaction (e.g., efficiency and task completion)?
- What are reasonable baselines and upper bounds, for example, learned fron human–human interaction?
- How do researchers conduct longitudinal studies and evaluation?
- What kinds of products are available that might make longitudinal evaluation (e.g., putting robots in people's house) possible?

While our focus in this chapter has been mainly on task learning where humans serve as teachers and robots serve as learners, it is not difficult to imagine that a well-trained and capable robot could also teach humans new tasks. In the intelligent tutoring world, computer programs have been teaching humans in various ways for more than three decades. Virtual agents teach humans all sorts of tasks, from turbine engine operation (Rickel and Johnson 2000) to negotiation (Gratch et al. 2015) to cross cultural communication (Johnson and Zaker 2012). The idea that robots might teach humans has received relatively little attention, perhaps in part due to the lack of capabilities. Robots are not yet teachers, but for many tasks (e.g., from doing experiments to manipulation of heavy equipment), the physical form of a robot will be useful in ways that computer programs and virtual agents are not. As robots become more capable, a reversal of the teacher/learner role is foreseeable and will bring further research challenges and opportunities.

From "Interactive Task Learning: Humans, Robots, and Agents Acquiring New Tasks through Natural Interactions,"
edited by K. A. Gluck and J. E. Laird. Strüngmann Forum Reports, vol. 26,
J. R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978-0-262-03882-9.