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.

7

Interaction for Task Instruction and Learning

Andrea L. Thomaz, Elena Lieven, Maya Cakmak,
Joyce Y. Chai, Simon Garrod, Wayne D. Gray,
Stephen C. Levinson, Ana Paiva, and Nele Russwinkel

Abstract

This chapter considers the qualities of human interaction and learning that will be most effective and natural to incorporate into any interactive task learning agent, and focuses specifically on the interactions involved in learning from explicit instruction. At the center of this interaction is a process that brings the common ground between a teacher agent and a learner agent into alignment. Errors or misalignments to this common ground drive the interactive learning process. The importance of timing is highlighted as is the dynamics of an interaction, as a communication channel itself, in this alignment process.

Introduction

What are the most effective and natural methods for humans, robots, and AI agents to interact in support of instruction and learning? To address this central question, we begin by establishing a context to frame the scope of our discussion, defining the landscape of tasks and learning interactions that we are considering for interactive task learning (ITL). We then introduce a model for ITL and discuss both the support for this model in natural human interaction and its implications for learning agents.

Group photos (top left to bottom right) Elena Lieven, Andrea Thomaz, Stephen Levinson, Joyce Chai, Simon Garrod, Ana Paiva, Nele Russwinkel, Stephen Levinson, Wayne Gray, Andrea Thomaz, Joyce Chai, Nele Russwinkel, Wayne Gray, Maya Cakmak, Elena Lieven, Simon Garrod, Ana Paiva, Maya Cakmak, Joyce Chai, Andrea Thomaz, Elena Lieven

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.

A question of contention raised throughout is: For ITL to be successful, is full human-level capability required? Our goal is to lay out the key features of *human–human* interaction and discuss ways in which principles of these interactions should be replicated in *human–agent* interactions. Importantly, these principles of interaction are implemented in a variety of ways with a variety of communication modes in human interaction. Thus, we expect that artificial agents may use any of a range of modes of communication to implement naturalistic communication principles.

Types of Tasks

In thinking about interactive tasks, it is useful to consider a "task space" that expresses gradients of difficulty on different scales. For instance, tasks can vary from a "simple" interaction with a physical domain (hammering in a nail), through an antagonistic interaction with an animate agent (zero-sum games), to a cooperative interaction where complex issues of joint control, synchronization, perspicuity of contributed actions, and plan reconstruction become especially crucial. Another dimension might be a hierarchy of modes of transmission of task knowledge between agents: from reverse engineering of a product, to emulation of an observed agent, through minimal (e.g., gestural) instruction, to full verbal and multimodal demonstration and instruction. A third dimension might be computational complexity, which is partly dependent on the other two hierarchies; for instance, cooperative interaction tends to be more complex than interaction with a physical domain because it also involves modeling another agent. Of course, physical tasks vary in their own complexity, too: picking up an egg and separating the yolk from the white, for instance, will require extreme delicacy of physical manipulation compared to picking up a ball and throwing it away. This three-dimensional space provides a way to think about the landscape of interactive tasks.

Consider the following different tasks, representing some of the variety encapsulated in this task landscape. The first is an example of a coordinative joint task: a healthcare robot that should help an elderly person who is not able to use his or her legs and needs help transferring in or out of a wheelchair. The actions to be learned (or the interaction with the person) should be flexible; for example, the action "lift the person up" would be quite different depending on whether the person is to be lifted up from a chair, from a bed, or from the floor. Here we need learning as a first step to train the robot to do a safe action and also to adjust the action according to the feedback of the person (e.g., "slower," "careful"). This interaction could take place via language (assuming the robot has some knowledge about what these adverbs mean) or by demonstration. The robot must not necessarily be able to talk, but it requires feedback mechanisms to give the person being cared for a secure feeling that the task will be carried out as expected.

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.

The second is an example of an agent system that could be an assistant system in a car that has to learn when to offer help or information to the driver. The system should also learn or anticipate when the driver is engaged in a task and should not be disturbed. The system should offer information when it is needed, at the appropriate time (e.g., "construction work on the road," "traffic jam on this route"). Here the agent needs to learn to model the driver and the tasks in which the driver is engaged, as well as the individual preferences of the driver. This system may need language processing to interpret the driver's language input and generate language feedback to the driver. This is not a cooperative task in the sense of equal partners; the agent system is providing support and should learn about the needs and goals of the driver through experience and feedback.

Although the agent systems from both examples share some aspects, teaching these two systems would be very different.

Another consideration is that tasks to be learned are often compositional and can be represented by, for example, and-or graphs (Liu et al. 2016) or hierarchical task networks (Mohan and Laird 2014; Mohseni-Kabir et al. 2015). An overall task can be broken down into subtasks (which can possibly be further reduced into other subtasks) with temporal and spatial constraints. A subtask can also be decomposed or implemented by primitive actions. Thus, learning a task will involve learning how to perform actions at different levels of abstraction, and this may require different forms of teaching. For example, a primitive action can perhaps be best taught through physical guidance, whereas a high-level task with partial order of subtasks may benefit most from language instructions (Chai et al., this volume).

Types of Interactive Learning

Let us now consider the types of learning found in humans and more precisely define which are most relevant to the ITL problem. Forms of interactive teaching and learning differ in two ways: (a) whether or not the modeler/teacher has an intention to teach and (b) how it is that the learner learns.

The action of a modeler/teacher may not consciously intend to teach but can still afford learning in the learner. One example in natural learning is chimpanzees learning to crack nuts: chimps crack nuts and young chimps stay close to their mothers, who tolerate "nut stealing," as follows:

Stage 1: Initially, the young chimp makes no efforts to crack the nuts but only to eat them once the mother has cracked them. This keeps the chimps near the mother and may lead to observational learning: the young chimps learn that *nuts can be cracked*. This form of learning is often called "stimulus enhancement" (Tennie et al. 2009).

Stage 2: The young chimp starts to bring nuts to the mother, demonstrating an understanding of the "goal" of the task (Boesch 2003).

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.

Stage 3: At a later stage the young chimp starts to try to crack nuts. Note that they do not coordinate the type of hammer, the type of anvil, or the nut type, and it takes many years of "trial-and-error learning" to achieve success. However, many conclude that "emulation" may be involved since the young chimp may have learned something about the kinds of actions needed as well as connecting this to the goal of these actions—the cracked nut (Tennie et al. 2009). They may, for instance, use the hammer/anvil that the mother has left. In this example, the issue is whether chimpanzee mothers are deliberately teaching. Boesch (2003) claims they are, whereas other researchers (e.g., Tennie et al. 2009) contest this interpretation. In terms of how learning from another agent takes place, Tomasello (1990) and others make the following distinctions:[1]

- Mimicry: copy the action; the goal is the action in itself.
- Emulation: copy the result of the action using other actions.
- Rational imitation: copy the modeler with an understanding of the intention behind the actions. For example, in Gergely et al. (2002), children are asked to imitate turning on a light when the modeler turns it on with her head, either (a) when she cannot use her arms because they are covered or (b) when her arms are free. Children turn the light on with their hands in the arms-covered condition ("she can't use her arms but I can"). They also turn on the light with their head when the modeler's arms are free ("she could use her arms but doesn't, so I should probably do the same").

Finally, natural pedagogy (Csibra and Gergely 2009) suggests a human-specific type of social learning through communication that speeds up learning (and avoids trial-and-error learning and statistical observational learning). This is a form of imitation in which the modeler presents a demonstration accompanied by cues that focus attention on specific elements, thus signaling that the cues are "intended" for the learner: slowing down, pointing, eye gaze, exaggerated actions.

Of course, not all learning takes place through interactions between agents, and each type of learning outlined below will have implications for how the learning is structured:

- *Entrenchment:* simply being immersed in the environment, the learner is exposed to experience from which to learn. For instance, children learn language through being entrenched in positive examples, not through explicit instruction.
- *Self-exploration:* the learner is on his/her own in the environment, learning through self-discovery and interactions with the world.

[1] Note: the precise definition and use of these terms is debated in the literature.

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.

- *Structured discovery:* something more like the self-learning that happens in a preschool, whereby the environment has been arranged to support particular kinds of self-learning.
- *Apprenticeship:* learning happens through emulation and imitation, by the learner doing the tasks that are modeled by an expert, but the expert does not necessarily have explicit instruction interactions with the learner.
- *Explicit instruction:* the teacher and learner enter into an explicit communication about the learner, with the joint goal of the learner obtaining some new task model. We are limiting ourselves to dyadic interactions, but this type of interaction can also happen in groups.

The learning interactions we consider most relevant for ITL are those with *explicit instruction*. When a human partner is interested in transferring task knowledge to an artificial agent, we expect this to be the most common form for that interaction to take. There may be contexts in which apprenticeship is appropriate, where a person wants to have their artificial agent learn by example without explicit teaching. For the purpose of this chapter, however, we focus on the scenario where explicit teaching is taking place.

Making the Task Learning Problem Interactive

Baseline Model: Two Agents plus World

Consider the following first approximation of ITL, shown in Figure 7.1 (see also Figure 2.1, Mitchell et al., this volume). For two agents, the learner (A_L) and the teacher (A_T), learning by A_L can be considered as the improvement of the performance metric (P) in the completion of a task through experience. Natural interaction between A_L and A_T shapes the experience, thus increasing the potential for interactive learning by A_L to occur. Note that both A_L and A_T are agents that have the ability to act in the world, W, and can thus change it

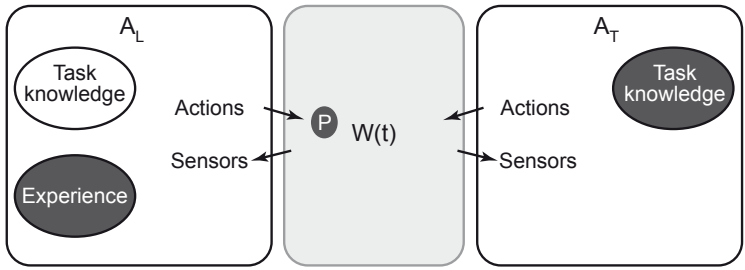

Figure 7.1 Baseline model for learning between two agents. A_L (the learner) and A_T (the teacher) interact in a shared world over time, W(t). Learning is considered to occur when A_L improves its performance (P) of a task based on experience.

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.

as well as each other. The learner perceives the world (and the communicative actions of A_T) as changes which occur in the world over time, W(t). A_T has the possibility to change the experience observed by A_L in a way that makes task learning more effective and efficient. Shaping the experience through the A_T's actions in this ITL framework aims to permit A_L to improve task performance. For that, communicative actions also function to keep the communication channel open and make the learner more engaged in the task learning.

Update to the Baseline Model

Many human joint activities involve what Bratman (1992) calls *shared cooperative activity* (SCA). He argues that SCAs depend on both agents recognizing that they are part of the SCA, and that they are committed to the goals of the SCA as well as to carrying out their part in the SCA. SCAs tend to have a clear lifetime with a beginning, middle, and an end. The cooperating agents may have to negotiate the beginning and the end of the SCA and thereby define the period over which their shared commitments apply.

Further developing the two-agents-in-world model of interactive learning, we propose the following changes to capture this commitment to the joint activity that is required for natural style interactions with a human teacher:

- Enrich the world state by including, for instance, more than just W(t).
- Enrich the representations within each agent.

In relation to the first change, enriching the world state, the knowledge and information that is shared between A_L and A_T should be explicitly represented; we refer to this as *common ground*. In Figure 7.2, we model common ground as a bulletin board that would minimally contain all the task-relevant information. Such a bulletin board would include the results of A_L and A_T's actions as well as any relevant props for their joint actions (e.g., furniture being constructed, bed and chair when lifting grandpa). In turn, we can distinguish between $W_{AL,}$ W_{AT} (corresponding to A_L's and A_T's own representations of the world) and $W_{AL1'}$, $W_{A2'}$ (corresponding to the common ground of A_L and A_T's knowledge of each other's world state). In general, joint activities are likely to be more successful when there is greater alignment between W and W′ (Sebanz et al. 2006). Thus, each agent has its own actions and perceptions of the world as well as a shared model of the state of the world that is built up between the two agents. There are several different types of information contained in this common ground representation. The agents need to maintain agreement/alignment on the task-relevant world state, the state of the learner's task knowledge, and the state of the interaction itself. Notice that ITL *requires* common ground, but also that the learning process itself *results* in new common ground through additional task knowledge.

In relation to the second change, enriching the individual agent's models, we need to differentiate actions into communicative and noncommunicative

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.

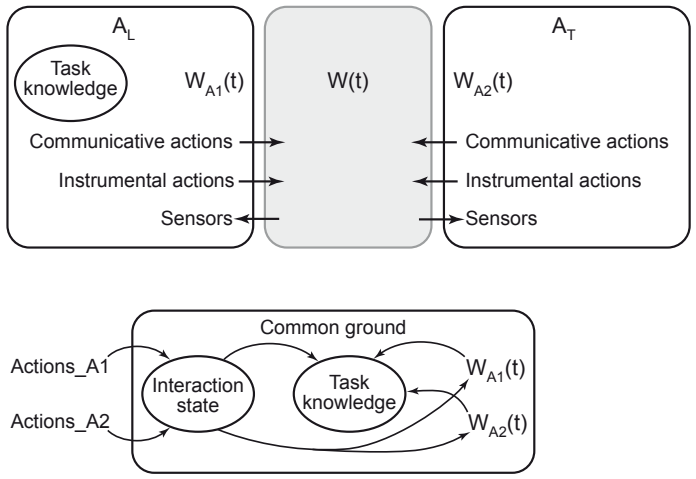

Figure 7.2 Enhanced model with explicit shared knowledge (common ground) between two agents, A_L (the learner) and A_T (the teacher), in a shared world over time, W(t).

actions (and perhaps communicative and noncommunicative sensing as well). More importantly, each agent could have mental representations of the information associated with planning the joint action. For instance, this could include information related to A_L and A_T's joint action planning (e.g., A_L needs to hold the pieces of furniture together for A_T to drill a hole for the screws). This could also include information about how A_L and A_T will implement their joint action (e.g., A_L holding the pieces of furniture at the same time and location as A_T drilling the hole). In general, joint activities are likely to be most effective when these representations are the same or aligned between A_L and A_T (Garrod and Pickering 2004).

Given this new formulation of the ITL agent model (Figure 7.2), we can now consider several different ways that the teacher agent can possibly shape the experience (E) of the learner:

- *Directing attention/referring*: this can be done by (a) manipulating the world to make some features, objects, or processes, which the teacher wants the learner to focus on, more salient; (b) using nonverbal communication (directing eyes, pointing) to the salient elements in the world (and in communication); and (c) by communicating verbally to focus attention. No matter what means of communication is used, referring has the result of bringing the two agents' common ground world state into closer alignment.
- *Manipulating the timing by which some actions occur* (communicative and environment changing actions): if a backchannel is communicated with some delay, for example, different meanings can be considered. These backchannels can provide signals like "keep going," "continue,"

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.

"give me more information," or "stop going on about this." This represents the dynamic nature of the interaction state, whereby time can implicitly move the interaction state forward. Nonaction must be considered an intentional action itself.

- *Providing a task description through verbal communication using natural language or some description language* (which at some extreme can be a programming language): this would have the goal of changing the state of task knowledge represented in common ground.
- *Demonstrating a task to the learner*: this requires synchronization between perceptions/actions/responses of the two agents and should involve alignment along all parts of common ground. The interaction state should have agreement on whose turn it is to act, the agents' world states should have agreement on what aspects of the perceptual space are relevant to the demonstration, and the demonstration itself should make a modification to the shared understanding of the task knowledge state.

We argue that this update to the learning agent model is what makes the learning process interactive, particularly when the interaction is with human partners. In the remainder of this chapter we describe ways in which human interaction is driven by the need to maintain these different aspects of common ground between agents. We first lay out more specifically what common ground means for an interaction, followed by how common ground is repaired when errors or misalignments arise. Thereafter we discuss how many important elements and characteristics of human interaction stem from the rather strict timing dynamics involved in maintaining common ground over an interaction. Finally, we close with a discussion of the multiple modalities humans use for such learning interactions and the extent to which agents need to have humanlike interaction abilities.

Common Ground

The "classical" version of common ground comes out of philosophy as a way of handling some of the ways in which languages distinguish material that is new to the interaction from older information that is already established (e.g., an oversimplified way of mapping onto the distinction between "a" and "the"). Common ground is normally considered to have different components:

- Basic facts taken for granted (e.g., we are in Germany, it's summer, gravity reigns on Earth, the capital of the U.S.A. is Washington, D.C.)
- The things in front of us that are mutually manifest (e.g., the books on the table in front of us both)
- The activity (or task) in which we are engaged
- How far we have progressed on that task

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.

- Tacit pacts about referring expressions (e.g., once we have referred to object X as "the O-ring" we should continue to use that expression)

Think about this as a blackboard: each time agents introduce entities and propositions, we add them to the list, so common ground is quintessentially incremental.

This classical version has problems. It assumes, for example, that everything interactant A believes, B should believe as well; there is no possibility of a quarrel or disagreement in such a world. Or, more relevantly for robotics, there is no individuation of what A and B know from what a new agent C knows. Thus, in such a learning environment, each agent must have its own "commitment slate": C's slate can be updated by interaction with A or B. For informational and skills learning tasks, the whole point of interaction is this updating, but the updating will happen in all cases: if we are moving a table together, we will serially update the location and orientation of the table.

In human–human interactions there are some categories of things that are assumed to be in common ground, such as naïve physics of the world/objects, referents which by default have space/time recency, actions that are usually rational (efficient with respect to goals), and other forms of common sense for humans. Agent designers should keep these aspects of human common ground in mind when considering what must be included in the innate abilities of an effective interactive agent. It would be great to have agents with these aspects of prior knowledge, but at the very least, agents should be able to communicate their limitations in background knowledge.

Developmental Trajectory of Grounding, Informing, and Referring

Babies start to take part in synchronized, dyadic exchanges early in life; there are arguments as to how early this happens, but definitely by two months of age (Feldman 2007). By around six months, dyadic games are well established (e.g., peekaboo, vocalization exchanges) (Rochat et al. 1999). At around nine months, infants show the beginnings of "intention reading/social cognition"; that is, the understanding that both they and others have communicative intentions (Tomasello 2008). This is signaled by episodes of joint attention in which the child and caregiver are sharing attention to the same object (Cameron-Faulkner et al. 2015) and making this manifest by gaze checking, mutual laughing, and pointing (Callaghan et al. 2011). Pointing has been much studied, and by about 12 months, babies point to share information, to inform, and to correct. Evidence for this comes from changes in the frequency/intensity of points dependent on caregiver response (Liszkowski et al. 2012). Tomasello (2003) stresses that the "gavagai" problem (Quine 1960) of referential ambiguity can only be solved through this system of intention reading, which allows the child to map the form of another's utterance to an inference about their

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.

intention in uttering it. Further, he argues that the rapid increase in word learning that occurs in the second year of life is dependent on intention reading.

In the early stages of language development, caregivers are usually the ones to provide common ground. Knowing what a toddler means can be very problematic without shared background knowledge and common ground. For instance, toddlers will use referring expressions like "it" without having previously used an identifying noun. But children start using determiners and pronouns in discourse fairly correctly (e.g., the use of "a" to introduce a referent and "the" for further mention) before the age of three. However, if usage is probed in experimental situations, it becomes clear that the subtleties of use (e.g., sensitivity to whether the interlocutor can or cannot see what is being referred to) take quite some time to develop (Matthews et al. 2006). If children are put into "standard" referential communication tasks where they have to identify a minimally different referent to a partner, their performance is surprisingly poor, both in terms of the identifying descriptions that they give and in terms of asking for clarification (Lloyd et al. 1995).

Grounding and Egocentricity

When a speaker refers to something as "the cup" (as opposed to "the big cup" or "the middle-sized cup"), this is supposed to indicate that there is only one cup in the common ground. However, when a speaker says "the middle-sized cup," this is supposed to indicate that s/he is referring to only one out of many possible cups in the common ground (i.e., the second largest of three cups). In other words, speaker and listener are assumed to take into account both the speaker's and the listener's perspectives on what they are looking at when making such references. However, there is much evidence to suggest that even skilled adult speakers and listeners do not always take common ground fully into account when referring to objects in a scene. Keysar et al. (2000) have carried out various experiments in which they arrange tableaux of objects, such that the speaker and listener have different views of those tableaux, and it is quite apparent to both that they have different views. For example, the speaker may see three cups (a big, middle-sized, and small one) whereas the listener can only see two (a middle-sized and a small cup). When describing the cup that is "middle-sized" from the speaker's perspective, which is actually the largest cup for the listener, the speaker will often refer to it as "the middle cup." In other words, the speaker will often refer from an egocentric perspective, not taking into account what is in common ground and available to both speaker and listener. Conversely, if the arrangement is the other way around—the speaker sees only two cups while the listener sees three—when the listener hears "the big cup," s/he will often look at what is the largest cup from his/her perspective, but not from the speaker's. In other words, the listener also tends to follow an egocentric perspective on what should be in common ground when s/he interprets the speaker's references. The effects are

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.

particularly pronounced when the speaker and listener are under time pressure, suggesting that common ground is not always fully taken into account even by skilled adult speakers and listeners.

This process is even more complex for robots and AI agents in that being physically present in the same space does not mean that humans and robots have the same perceptual access to the shared environment. A robot has much different perceptual, motor, and reasoning capabilities than a human. The robot's representation of the shared world is significantly misaligned from the human's. The lack of a joint representation makes grounding between humans and robots extremely challenging, yet essential for the success of an interaction and a baseline from which learning can take place. As shown by Chai et al. (2016), humans and robots will need to make extra effort to bridge the gap and strive for a common ground of shared representations.

In the context of ITL, it is reasonable to assume that humans may be better than robots at detecting and remedying missing common ground. Thus, it is important for the robot to take extra effort to provide sufficient cues to assist the human in detecting and repairing missing common ground in a timely fashion. One potential device is to make the robot's internal representations transparent to the human through, for example, language description or visual display (Alexandrova et al. 2014; Hayes and Shah 2017). Another device is through confirmation, commonly employed in dialogue to help establish common ground. There are two types of confirmation: (a) explicit confirmation, where an agent always explicitly asks for a confirmation about its understanding (e.g., "you are talking about this cup, correct?"), and (b) implicit confirmation, where the agent provides an implicit confirmation of understanding, as through the relevant actions on their next turn (Litman and Pan 2002). There are trade-offs between explicit and implicit confirmations. Explicit confirmations make it easier for humans to correct mistakes and can lead to better task success rate, but they are cumbersome and can result in lengthy interactions. Implicit confirmations are much more natural and quick, but there is risk in the delay of detecting and repairing mistakes. In human–robot referential communication, studies have shown that the robot's belief of the reference (referred to by the human) may often not be the same as the reference intended by the human (Chai et al. 2016). In this case, if the robot replies a generic "got it," this may fool the human into believing that common ground has been established when in fact it has not. This simple acceptance from the robot is more detrimental to common ground than a simple rejection ("I don't get it"). When the robot provides information about its internal representation of the believed reference (e.g., through language descriptions), common ground can be significantly improved. Thus, it is important for the agent to adapt different types of confirmation under different situations.

In addition to handling explicit communication about common ground, agents need to account for the fact that much of the common ground between humans is inferred implicitly rather than explicitly discussed, as in the

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.

goal-directed action inference example mentioned previously. When children see a woman hitting the light with her head when her arms are free, thus indicating that the use of the head is intended since her hands could have been used, they interpret this as an instrumental action (hitting the light) *and* a non-instrumental action (not using the most efficient means available), which implicitly communicates that the goal of the task *includes* the action of hitting with the head, not just turning the light on (Gergely et al. 2002).

Bringing the perceptual and action capabilities of robots or AI systems to human levels will certainly help with most of the common ground issues and could enable interactions at the level of human–human interactions. However, realizing this is extremely challenging. Alternatively, if robots had all the capabilities necessary to perform the range of tasks they might need to learn and there was broad acceptance and usage of best practices for ITL implementations across robots and AI agents, humans might get used to the robot's limitations and still interact with them smoothly. This is similar to how humans interact with pets, where the expectations are lower but accurate.

Humanlike transparency mechanisms (e.g., gaze, pointing, head gesturing) can exploit people's ability to interpret mechanisms without any instruction or training. Implementing those mechanisms with precision in robots, however, is difficult. Also, robots performing these actions might bring an additional expectation that the robot can also process such information from the human, which is even more challenging. Instead, visual transparency channels on a robot, such as a screen or projection from the robot onto the environment (which are not afforded in human–human interactions), can afford high bandwidth visual information transfer and might be effective in human–machine interactions.

Repairing Misalignments in Common Ground

Next let us now turn our attention to the concept of "errors" in common ground. In a sense, this is the entire reason for a learning interaction at all: to repair the misalignment in common ground between a teacher and a learner. Errors trigger the requirement for an ITL process. Without errors there would be no need for ITL. Continually throughout an interaction, misalignments in all aspects of common ground are being detected, diagnosed, and repaired.

Detection happens through self-monitoring (broken expectations in perceptions or actions), implicit feedback (back channel), and explicit feedback ("stop, that was wrong"). There are many different sources of error to be detected and diagnosed:

- *Perception errors* occur during perception (e.g., the agent fails to perceive an object in the world).
- *Task execution errors* are caused by faulty execution of an action in the world (e.g., the robot does not reach an object in the world).

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.

- *Representation errors* encompass misconceptions as well as incomplete or incorrect task representations. This is related to the Brown and VanLehn (1980) generative theory of human misconceptions, or bugs, which can be seen as resulting from procedural skills acting on incomplete or incorrect procedures (i.e., tasks).
- *Communication errors* result from a lack of common ground or inadequate mechanisms (e.g., feedback) to sustain the flow of communication with the teacher agent.
- *Anticipation errors* occur in the prediction mechanisms of the agent and lead to expectations (future state of the world) that are not possible to meet, given the available actions and resources.

Once an error is detected, it can be self-diagnosed and repaired or the diagnosis and repair can happen through subsequent interaction. In human conversation, self-corrected repairs can be nearly immediate (within ca. 700 msec, at a rate of about once every 80 sec; Dingemanse et al. 2015). But situations that Norman (1981) refers to include longer time-frame errors, as do many that Reason (1990) discusses.

Human Interaction Is Built to Minimize Errors

When repairing errors or misunderstandings in dialogue, effort is distributed between the two interlocutors. In the following telephone conversation (from Drew 1997), for instance, the listener helps the speaker detect the source of confusion (the word "gorillas") by interrupting with "forty-nine what?"

Hal: .an‘ Leslie ’t was marv’lous (.) D’you know he had (.) forty nine [?]
g’rillas. .hh th-there. (b) (.) br[eeding in ( ) [?]
Lesley: [pf- f- Forty nine what? [?]
Hal: G’rillas. [?]

Dingemanse et al. (2015) argue that the effort of identifying the source of the problem to be repaired and repairing it is nicely distributed between the two interlocutors to minimize the time spent repairing the dialogue.

In thinking about how errors in robot–human interaction could be handled, it is useful to refer to the human–human system and to try to extract general principles that might be useful. Let us take the language user and consider the production or execution system. The person starts out with an intention, recodes this into a semantic specification, which in turns gets recoded into a syntactic specification, which is then fleshed out in an abstract sound system (phonology), which in turn is recoded into articulatory (muscular) instructions. At every representational level, the human system probably does error checking (e.g., legal expression checking, checking the derivation form for each of the prior levels). What we know with certainty is that there are two self-monitoring levels: a so-called "inner loop," which is prearticulatory, and an

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.

"outer loop," where as one says something, one checks that it corresponds to the intended sequence. In many cases, the inner loop can catch errors before there is any overt sign at all. At other times, there may be some mild perturbation (e.g., a pause or vocalization, "um") in the overt signal. Where the error is detected by the external loop, the speaker interrupts herself with an audible glottal closure, signaling "oops," and then recycles the earlier delivered chunk back to the point of the error: "You go left at the corn… you go RIGHT at the corner." At this point the speaker may think she has completed her turn, but the interlocutor may now miss the beat (ca. 200 msec after the turn ends), at which point a response is expected (perhaps also leaning forward or frowning), indicating some possible hitch in comprehension. This provides a space (ca. 300 msec) for the error speaker to self-repair or augment, "You go left at the corn… you go right at the corner, (500 msec) at the intersection." If the speaker misses this opportunity, the addressee in difficulty can still trigger (initiate) self-repair by the original speaker:

A: "You go left at the corn… you go right at the corner, (500 msec) at the intersection."
B: "The intersection of Bryant and East Street?"
A: "Yes, by the 7Eleven."

Sometimes, the misunderstanding only becomes apparent later:

A: Do you know who's going to the meeting?
B: No, who?
A: I don't know…

Here the response to B's turn displays a misunderstanding that A's turn was a preliminary to a telling, when in fact, as turn three makes clear, it was a simple question. This illustrates the utility of a communication system that alternates short turns across speakers; the responses indicate whether the prior turns were understood as intended (Sacks et al. 1974).

By looking at the whole system of incremental possibilities of repair, one sees that the whole interaction system is designed to give multiple successive opportunities to catch misunderstandings and errors (Schegloff et al. 1977). The system is optimized for efficiency, first within the speaker's self-monitoring loops, then through overt self-repair (where the speaker foresees errors or upcoming misunderstandings), then through a pause inviting self-repair, and finally (and reluctantly) through other-initiation of repair. The latter involves an inserted, potentially disruptive sequence. The disruption is minimized by an ordered preference of types of other initiation: the listener that does not understand gives a precise localization of the problem ("the intersection of Bryant and East Street?" is more efficient than "Where?"). Interestingly, the sum of the length of initiator and repair tends to be no longer than the original troublesome utterance (Dingemanse et al. 2015). One reason for this efficiency is that other-initiation of repair occurs every 80 sec in natural conversation.

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.

Learning from Mistakes

What does it mean to be wrong? What do errors tell you? One explanation is that what you did was wrong and requires follow-up; however, there are several learning paradigms in which errors are integral parts of the learning process (Lorenzet et al. 2005). There is utility in exploration. Importantly, though, there must be bounds beyond which the learner knows it cannot be allowed to go, based on safety or cost (or other criteria). The role of errors may be different for humans and agents. For example, children may need to make an error to learn effectively from it. However, a robot may have a learning mechanism that allows it to learn based on communication that substitutes for the experience of actually making the error.

Children sometimes make errors on purpose, to gain a better understanding about the consequences of some action and better predict the future. If an action and its consequence are not known by the child, the child will be curious to explore it. What happens if the glass falls? Does it break? Does it not break? What happens after a glass has broken? Exploring this gives the child a more complete picture of the world.

For robots it is also important to explore the environment and to explore what works and what does not (within a certain range, of course). If a glass is grabbed too hard, it will break; if the door handle is turned too gently, it will not move. Exploring the range of different outcomes from an action gives a better and richer representation to operate more flexibly.

For these kinds of "errors," which function more as exploration actions, the robot/child will probably need to realize on their own when an error occurs. For some types of errors, this might not be obvious to the robot, thus necessitating the support of the teacher. Support is also needed if the reason for an error is unknown to an agent. In the end, it is crucial to know what action would *not* cause the error.

Controlling and guiding this exploration process is an important role that the human teacher can play for the learner; that is, helping the learner collect the most informative "near miss" examples that will lead to generalization.

Learning systems can end up in a sort of good enough, less than optimal state (e.g., Klein and Perdue 1997). Consider a landscape of solutions where there are many, locally optimal, sort of OK solutions, but only one or two optimal solutions. For a machine to escape from its less than optimal cul-de-sac, it will need to be nudged out of its local optimum and forced to explore the larger landscape of possibilities. This can be achieved, as in Bayesian modeling of fitness landscapes, by perturbing the current state: the system is forced to go downhill for a bit and then starts to climb another incline that may turn out to be the global maximum (e.g., Markov chain, Monte Carlo, Metropolis coupling in Bayesian phylogenetics, as in Reesink et al. 2009). Gray and Lindstedt (2017) talk about "plateaus, dips, and leaps" (see also Gray et al., this volume). "Plateaus" contrast with "asymptotes" (Gray 2017). Asymptotes

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.

reflect performance at a theoretical limit whereas plateaus are periods of stable but suboptimal performance. Better methods can yield better performance but the agent may not have knowledge of the better method or possess the skills required to master the method, or simply not care to get better.

Different teaching styles result in different ways that a misalignment on task knowledge is handled. Consider the following task instruction scenarios:

- Marie is a preschool teacher who wants to teach her students how to mix colors to obtain other colors. Instead of just telling them the different combinations, she uses a technique called "provocation." She presents the students with an uncolored picture of a frog and two tubes of color: blue and yellow. She says "this is all we have, what should we do?" Some students will give up and say that it is not possible; others will paint the frog blue or yellow (which is not considered a mistake); a few will try something new and mix the colors to obtain green. Children who go through this discovery process are much more likely to remember the outcome than those who are just told about it or observe it (Craft 2001).
- Kenan is a carpenter certified to teach woodworking. He regularly works with apprentices at his workshop. When apprentices start working at the shop, Kenan first assesses skill levels by having them perform basic actions, such as cutting with a saw or sanding. He gives them a task like "cut all this wood into 16 cm pieces," commensurate with skill level. When teaching a new task, Kenan uses a supervised discovery process. He tells the apprentice to attempt the task and interrupts them when they go wrong. For example, he may tell an apprentice, who has never previously done the task, to glue two pieces of wood together. If the apprentice goes down a wrong path, such as starting to apply glue before making certain that the surfaces to be glued have been sanded completely flat, he then interrupts the apprentice's work to bring him/her back to the correct path. When teaching how to use more dangerous cutting and milling machines, Kenan first demonstrates the use of the machine and then tightly supervises apprentices as they try to use the machine.

One key difference between these scenarios is the extent to which mistakes by the learner are allowed, or even encouraged, as a function of situational characteristics. The same provocation-based discovery learning process that is entirely appropriate in the context of interactions meant to improve understanding of color may be inadvisable in the context of interactions meant to improve woodworking knowledge and skill. Kenan cannot afford to let his apprentices explore possible paths that might lead to waste of expensive resources and, perhaps even more importantly, an increase in risk of harm. The woodworker's blade is less forgiving than the artist's brush. The difference between these two scenarios may greatly impact the role of exploration needed

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.

for ITL in robots and AI agents. It also emphasizes the important role of situation understanding and contextual reasoning in selecting modes of interaction.

Interaction Timing and Synchronization

So far we have focused primarily on the substantive context and content of interaction, leaving implicit in the discussion an inherent characteristic of all interaction: it progresses over time. Timing of an activity is separate from content. In this section we detail the importance of timing as a first principle of the interaction.

Correct timing and synchronization are crucial in many aspects of human interaction. The turn-taking system in conversation operates with ca. 200 msec turnaround (Stivers et al. 2009), the normal human minimum response time for the simplest preplanned response. This is far too short for planning spoken utterances, which generally require at least 1 sec (600 msec for a single word, 1500 msec for a simple clause) before output can begin (see Levinson, this volume). This implies that speakers are predicting the ends of the incoming turn and planning their own so that they are ready to respond on time. The speed may have origins in the phylogeny of our communication system, before the complexity of linguistic signals developed (Levinson 2016), but it is also maintained by the semiotics of delay. For example, neuroimaging has shown that as a gap after a prior turn lengthens, expectations change: we usually formulate questions to favor a *yes* answer, which is expected at the normal ca. 200 msec response time; if the answer *no* comes in at the normal response time, it evokes an N400 or surprisal reaction, but this evaporates over time as *no* becomes more probable (Bögels et al. 2015a).

One issue for timing of interactions, especially when a machine is teaching a human, is that human tutees often interpret a slight pause before positive or neutral feedback as signaling negative feedback (Fox 1991, 1993). For instance, if a student answers a tutor's question and then gets a slight pause before hearing "yes" or "umm," the student will often infer that the answer is incorrect.

In general a delay in response after a response-requiring turn signals that an unwelcome response is likely. Withholding response after any turn can signal that its import was not clear. In general, then, timing is part of the signal in human interactions. This may be very problematic for human–machine interaction, but it is worth noting that there are classes of humans, most notably children, who may be much slower than the human norm, and adults are pretty good at adjusting expectations to childhood norms. It may be better for ITL systems to signal their timing limitations (perhaps by junior stature) than to attempt to meet full normal human speed of response.

Synchronization is a low-level, probably largely unconscious, process whereby two agents come to coordinate. The simplest case can be modeled as

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.

coupled oscillators, as when the seventeenth-century Dutch scientist Christiaan Huygens noted that two pendulums mounted to the same structure come to synchronize over time. In biological systems, quite complex behaviors, like the synchronized firing of a swarm of fireflies, arise in a similar way (here, by resetting a biological capacitor when the neighboring firefly turns on). Humans playing music together, for instance, tend to harmonize brain oscillations, thus providing a shared internal metronome. Finely timed coordination may well depend on this, but for various reasons it will rarely be sufficient: fine timing may also depend on predicting the other's action culmination and even preplanning one's own productions so they are ready to go. Coordination is prediction plus generation in a joint activity, requiring mental simulation of the other.

Human speech communication has the distinctive property of alternating communication bursts between speakers. During an incoming utterance, an addressee may be signaled out by gaze and normally gives feedback signals at major chunks of incoming material. This becomes especially obvious if speaker A is delivering a story which, given the turn-taking structure, is often implicitly negotiated at the beginning:

- A: "Did you hear what happened to Joan?"
- B: "NO, what?"
- A: [chunk1], [chunk 2], [chunk 3].

In overlap with the end of chunk 1, B is likely to say "mm"/"uhhuh" or the like, or nod, thereby recognizing (a) that chunk 1 has been received, (b) there is nothing in chunk 1 that is causing comprehension difficulty, and (c) that B has nothing compelling to say at that point. The opportunity to do this recycles at the end of chunk 2, 3, and so on (Schegloff 1982). The class of relatively content-free back channel responses is fairly limited per language, with upgraded versions also available (e.g., surprise markers "wow," empathy markers "oh dear"). These signals do not count as turns, which is why they typically occur in overlap. Conversely, longer phrases (less limited of course) are likely to count as turn initiators, leading to expectations of possible speaker switch. Back channels of this kind, as the name suggests, thus essentially signal "channel open, message received," while implying by virtue of the activities that were not done instead (e.g., initiation of repair or a new major response) that full understanding has occurred. B recognizes that A is producing a longer stretch of speech that has not finished.

Through a back channel, a listener demonstrates to a speaker his/her continued interest in communication. As back channels play an important role in coordinating human–human conversation, it becomes important for agents to have a capability to generate back channels during human–agent communication. This involves making a decision on when to generate back channels (i.e., the timing of back channels). In human–human communication, back channels occur very fast and seem to be elicited by the speaker based on a variety of

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.

prosodic, verbal, and nonverbal cues (Schroder et al. 2012). Previous works in conversational virtual agents have developed predictive models (e.g., sequential probabilistic models) to predict the timing of where a back channel should be generated (Morency et al. 2008). As there is evidence that human listeners generate back channels even without attending to the content of communication, these previous predictive models often only consider surface features, such as prosody, pause, gaze, and direction from the speaker. Their empirical results have demonstrated that generating back channels that are synchronized in time with speaker contributions is an extremely challenging task.

The nature of communication in ITL as well as in conversational virtual agents (e.g., in the context of social communication for negotiation, consultation, and therapy) are quite different. It is not clear whether previous work on virtual agents can be directly applied to ITL. The prediction may no longer depend on acoustic but rather visual features (e.g., observed from human demonstrations). In ITL, the chance that an agent might misunderstand task instructions given by a human (whether verbally or through demonstrations) is high. Since humans may perceive back channels as an explicit confirmation of understanding, generating back channels will need to be tightly linked to content processing (rather than surface cues). This could delay the appropriate timing for generation. Without connecting to content processing, a back channel may have a danger of leading the speaker to believe a task instruction has been successfully understood and then later discover otherwise, causing a high cost repair for the downstream communication. Thus, when to generate appropriate back channels in ITL remains a challenging research question.

Chao and Thomaz (2013) have shown the importance of timing control in human–robot interaction and its social impact. Their work on the CADENCE system shows that manipulating these turn-taking timing parameters (e.g., space between acts, likelihood of interrupting the partner) results in robot behavior that people perceive as being significantly different. Moreover, people attribute different personalities to the robot; changing the robot's personality by manipulating these timing parameters results in different behavior from the human partner, thus manipulating the social dynamics of the dyad.

Conclusions

In this chapter we have considered the qualities of human interaction and learning that will be most effective and natural to incorporate into any ITL agent, specifically focusing on the interactions around learning from explicit instruction. We argue that this type of learning is centered around bringing the common ground between these two agents (teacher and learner) into alignment. Thus, errors drive the ITL process by triggering interaction and learning. Without misalignment of common ground, either through errors or missing knowledge, there would be no need for ITL. Finally, we highlight the

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.

importance of considering timing and the dynamics of an interaction as a communication channel itself, as well as emphasize the importance of analyzing the extent of shared capabilities between the two agents.

What all of this argues for in ITL with robotic and AI agents, is that information from multiple short turns of interaction between the teacher and learner will have the best opportunity for minimizing errors in communication that will arise naturally. Short bursts of information between the two interacting partners is likely to be the most successful way to transfer task knowledge between the two, incrementally updating errors in common ground until the teacher and learner come into alignment.

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.