16

Early Developing Prerequisites for Human Interactive Task Learning

Franklin Chang

Abstract

Humans are better than artificial computational systems at learning to do new tasks through interaction. Part of this ability stems from preexisting capabilities that appear early in human development. Children have internal physical models of how objects move and they attribute mental states (e.g., goals, beliefs) to objects when their behavior is unpredictable. They are also able to develop context-specific rules and identify how to help others achieve their goals. To explore how these abilities can be transferred to interactive task learning (ITL) systems, this chapter proposes a world-state prediction model. The prediction model can learn detailed physical regularities in the environment and is able to develop representations for predicting the actions and goals of animate agents. The model suggests that prediction and prediction error are capabilities that could improve ITL systems.

Introduction

Humans are good at learning new tasks through interaction. If you show people how to play a new game, where one needs to find monsters in different locations in the real world and throw objects at them using your smart phone screen, humans can easily learn to play that game without extensive training. This ability to learn a new game through interaction with a human teacher differs greatly from the learning ability of existing artificial agents. Recently, a deep learning neural network was shown to be able to learn to play 49 Atari video games by mapping between pixel-based images and the joystick actions needed to play the game (Mnih et al. 2015). Although the ability of this neural network to learn to play these games is impressive, each game requires extensive training based on a massive database of games. Also, while human experts

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.

are good at playing new games (Green and Bavelier 2012), these deep learning systems showed a lower accuracy when learning multiple games within the same system (Kirkpatrick et al. 2017). Part of the difference between human and artificial systems arises from abilities that appear early in human development, which shape the way that learners understand new tasks. Using studies of infants and toddlers to understand these early abilities, I will attempt to link these abilities to interactive task learning (ITL).

Although most people assume that infants exist in a "great blooming, buzzing confusion" (James 1890), research with infants has revealed that they appear to understand a range of constraints on the physics of objects (e.g., occlusion, support, collision, containment; Baillargeon and Wang 2002) and their interaction with other moving objects (e.g., Leslie and Keeble 1987; Frankenhuis et al. 2013). In addition, infants have social knowledge about how animate entities differ from inanimate entities in their motivations for action (e.g., goal-directed motion; Woodward 1998; Luo and Baillargeon 2010). Since these infants were not explicitly trained to understand the novel scenes used in these studies, this research suggests that their event understanding systems may have features that enhance their ability to automatically understand both physical motion and goal-directed action. These abilities give humans a rich predictive representation of their world even before they begin to interact with a teacher in learning a particular task.

Early Developing Abilities in Human Children

We cannot ask preverbal infants about their understanding of the world. Our knowledge about their abilities comes from two tasks that measure their expectations: habituation and violation of expectations paradigms. In the habituation paradigms (Spelke et al. 1992), infants are exposed to a scene multiple times until they habituate or become bored with the scene; they will often look away from the scene when this happens. When a new scene A is shown and the child is still bored or uninterested, this is seen as evidence that the child views scene A as being the same type as those that were viewed previously (even if it is different in some way). If the child becomes interested upon viewing a new scene B (dishabituation), this is taken as evidence that they view scene B as being distinctly different in an important way. If scenes A and B differ in some particular dimension, then dishabituation provides evidence that infants have knowledge of this dimension. In violation of expectation paradigms (Baillargeon and Wang 2002), children are shown an event that conforms to their expectations. Then they are shown an unexpected event that violates these expectations. If children look at the unexpected event longer than the expected event, this indicates that they have some knowledge about the features that differentiate these two events. Habituation involves learning about the event within the experiment, whereas violation of expectation involves knowledge gained outside 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.

experiment. These two types of paradigms allow us to probe the knowledge that young children have about visual events.

The Atari deep learning system started with pixels and had to learn the entities associated with different games (e.g., aliens, frogs). Although the human vision system also has a pixel-like input in the rod and cone cells of the eye, the brain seems to process the world in terms of objects that can exist even when they are not visible. Baillargeon et al. (1985) showed that seven-month-old infants were surprised when an object was placed behind a screen and the screen was rotated such that the object seemed to have disappeared. This violation of expectations study demonstrated that infants believe that objects take up space (spatial extent). Objects are also collections of elements that move as connected and bounded wholes. Spelke (1990) found that if two objects move relative to each other, even though they were constantly attached or connected, infants perceive them as two objects. Furthermore there is evidence that infants expect objects to move on connected, unobstructed paths and are surprised when objects appear to move invisibly across space or through other objects (Spelke et al. 1992; Aguiar and Baillargeon 1999). Therefore, within the first year of life, infant understanding of the world is not based on a raw list of pixel values, but rather on an internal model of objects with constraints on how they move and their spatial extent.

If representations are object based for infants, then infants must be tracking objects as they move around the scene. Leslie et al. (1998) have argued that infants and adults use one system for object tracking, which involves a set of pointers that stick to a particular object (dorsal *where* system), and another system to encode the visual features of each object (ventral *what* system). This is different from the approach in the deep learning Atari system, which utilized a series of convolutional neural networks to mimic the gradual abstraction of features in the ventral part of the visual system in the human brain, but did not have an explicit system for object tracking (e.g., a single frog on different parts of the screen are treated as different "objects"). Evidence for these visual pointers comes from multiple object tracking studies, where adults see videos with multiple identical circles moving around randomly (Pylyshyn and Storm 1988). Specific (white) circles are identified as targets (e.g., colored in red) and then the circles are made identical again (e.g., changed back to white). The circles then move around randomly while participants stare at a cross in the center of the screen. Later, participants are queried about a single circle and must say whether it is a target or not. Due to the fact that the circles are identical during the random motion, there are no visual features that can be used to track the circles, so the only way to identify the targets is if the participants are following them through the whole trial. These studies suggest that participants can track a limited set of objects in parallel using a set of visual pointers.

Support for these pointers in infants comes from a study by Spelke et al. (1995), where infants viewed two screens, separated by a gap, and saw an object move behind the first screen and another object (with the same shape)

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.

emerge later from behind the second screen (Figure 16.1). In the discontinuous condition, it appears that there must be two objects, because an object cannot cross the gap without being visible. In the continuous condition, the object appeared in the gap between the two screens such that it looked like one object was moving behind both screens. Infants were tested with similar scenes without the screens and they showed that they preferred the test scene that matched the training scene. If the infant assigned a pointer to the first object, then the same pointer could be used when it reappeared in the gap as well as when it appeared after passing behind the second screen. In the discontinuous condition, the appearance of the object from the second screen would require a new pointer. Additional studies by Xu and Carey (1996) showed 12-month-old infants scenes where two objects with different shapes appeared from behind a screen one at a time. At test, the screen was lifted and they were shown an expected scene with two different objects or an unexpected scene with two objects of the same shape, and they were surprised by the unexpected scene. These results suggest that infants track objects in scenes, even when they are not visible, and their representation of scenes involve these object-based representations. Object tracking is critical in any interactive task where an agent needs to communicate about multiple identical objects. For example, to learn how to barbecue multiple similar-looking items on a grill, a learner needs to track which items have already been flipped by the teacher as well as which items still need to be flipped (as well as any that might have accidentally fallen on the floor in the flipping process). Thus, the work with infants suggests that humans have a task-independent multiple-object tracking ability that could help support these types of interactive tasks.

After an infant can track objects, it becomes possible to identify the nature of the resulting interaction. One study that provides an important insight into

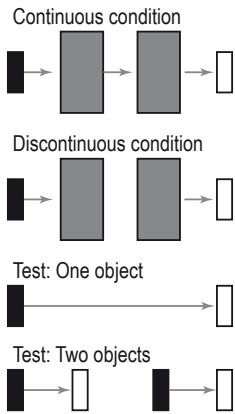

Figure 16.1 Two conditions in which object tracking was tested in 12-month-old infants (Spelke et al. (1995). Arrows indicate the path of an object (black box) as it moves behind two screens (large gray box) to its final position (white box).

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.

how this takes place was conducted by Gao et al. (2009), who used a multiple object tracking study to examine the features that support identifying a chasing action. In this study, multiple identical circles moved in a random manner. One of these circles was the wolf (the rest were sheep) and it had the property that it was always moving toward one of the sheep (central circle is wolf in Figure 16.2). Gao et al. found that adults were better at identifying the wolf when the wolf's angle of motion toward the sheep was more direct ("heat seeking"). Because the wolf and the sheep were all identical circles, participants had to first track the objects using pointers and then encode visual heuristics between pairs of pointers, such as the directness of the angle of motion. These visual heuristics appear early in development, as four-month-old infants prefer videos with chasing as opposed to those without chasing (Frankenhuis et al. 2013). This shows that children and adults are not just tracking objects, but associating relational features with each object pointer, such as angle of motion relative to other objects in the scene.

When chasing is taking place, the "wolf" consistently tries to move in the direction of the "sheep." In other actions, there is more temporal structure to the interaction. One event that has been extensively studied involves pushing or launching actions (Michotte 1963), such as when a green ball hits a red ball and pushes it away. This event begins with the pusher (e.g., a green ball) moving toward the pushee (e.g., a red ball). Critically, the pusher should make contact with the pushee, and the resulting motion from the pushee should begin without delay after contact. An understanding of the effects of these constraints on causal actions is evident early in acquisition. Leslie (1984) habituated six-month-old infants to videos and looked at how much they dishabituate to reversed versions of videos. Results show that they view the reversed one as

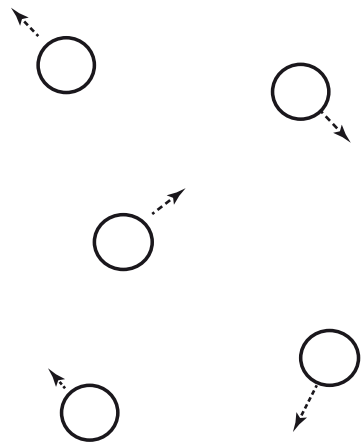

Figure 16.2 Schematic of visual heuristics used by adults to track multiple objects in a chasing action: the "sheep" are moving away from each other while being pursued by the "wolf" (central circle) Adapted from Gao et al. (2009).

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.

being substantially different than the one to which they were habituated. They found that reversals of pushing actions yielded more dishabituation than reversals of videos with a single object moving across the screen. They also found that infants were less likely to dishabituate when the video had no contact between the pusher and pushee or when there was a delay in the movement of the pushee (Leslie and Keeble 1987). This work suggests that infants have multiple innate features/heuristics that they are able to combine in various ways to recognize causality. When observing an item knocked off of a grill, during the course of flipping other items, a person can identify the cooking tongs (used for flipping) as the cause of the action (item falling on the floor) without prior training. Thus humans have a range of task-independent features that are applied to multiple objects in parallel and which combine to give them a better understanding of the causal structure of real world events.

One account of how humans understand these actions is to assume that humans take a *teleological stance* (Gergely and Csibra 2003), where actions are perceived as a rational means to achieve a goal state under certain situational constraints. Evidence in support of this position comes from studies that use videos where a ball jumps over a wall; this suggests that the ball has the goal of getting to the other side (Figure 16.3). Later, when the wall is removed, one-year-old infants were surprised when the ball jumps along the same path, because it could have moved in a direct path toward its goal. On this view, infants are able to identify the goal state of the ball and how the situational constraints of the wall block the direct motion toward the goal (Csibra et al. 1999). Thus, the most rational approach to achieve the goal would be to jump over the wall when there is a wall, and to go straight when there is no wall. This ability to

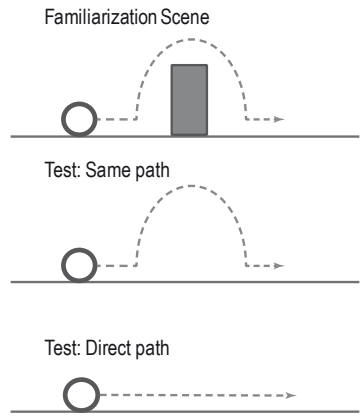

Figure 16.3 Schematic depiction of the teleological stance. During familiarization, one-year-old infants observed a ball jumping over a hurdle to reach a goal state (top). Two test states followed where the ball followed the same path (middle) despite the hurdle being removed and where it followed a direct path (bottom). The infants were able to identify the situational constraints. Adapted from Csibra et al. (1999).

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.

see how action is dependent on situational constraints is an important ability for ITL systems. In the grill example, if a teacher was trying to flip item A and accidentally caused item B to fall off of the grill, the human learner would recognize that the arm motion of the teacher was not the most rational/direct means of causing item B to end up on the ground, and thus the teacher's actual goal must have been to flip item A.

Up to now, we have looked mainly at studies that involve whole objects without any moving parts. But it has also been shown that people can recognize biological motion in point-light displays (Johansson 1973). In these displays, performers with multiple lights attached to their body at various points (e.g., hand, legs, head) perform some action (e.g., running, jumping) in a dark room, where only the lights are visible (Figure 16.4 shows two frames from a running video). When these videos are shown to adults, they are able to label the action that is being depicted (Johansson 1976). Golinkoff et al. (2002) also showed two point-light displays of different actions (dancing, walking) to three-year-old children and found that when the children heard a verb that matched one of the actions ("look at dancing"), they turned their head toward the appropriate video. Since the mapping of specific actions and words must be learned from experience, these abilities appear after three years of age, but the ability to understand these videos emerges earlier: four- to six-month-old infants exhibit a preference for a point-light human walker over an inverted walker or random motion (Fox and McDaniel 1982). Bidet-Ildei et al. (2014) found that even three-day-old infants prefer walkers over random motion, even when there was no horizontal translation across the screen (as in normal walking). These studies demonstrate that the infant mind is ready to understand biological motion from birth. One way to explain this ability is that the mind attempts to predict the motion of the points in these displays and biological motion is more predictable, because there are correlations between the motion of different points in these displays. This work suggests that human understanding of action does

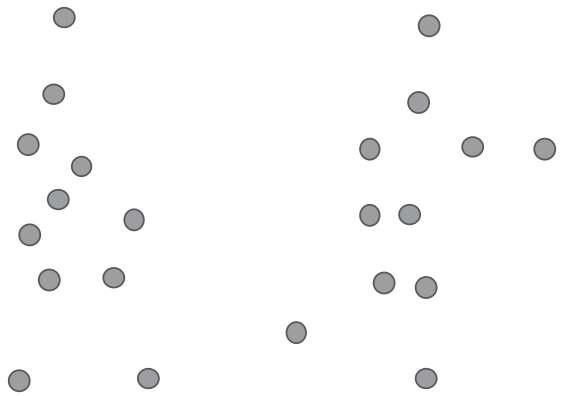

Figure 16.4 Two point-light displays of a man running.

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.

not involve learning rigid action rules; instead, predictive learning mechanisms capture the complex motion patterns present in biological systems.

To explain how we process biological motion in point-light displays, Giese and Poggio (2003) proposed a computational model that used the fact that the brain has distinct pathways to process form and motion. The ventral *what* pathway is specialized for form information and it starts initially with small receptive fields, which focus on small features (e.g., the orientation of lines), and gradually expands to larger receptive fields, which eventually span the whole object. The dorsal *where* pathway is specialized for motion information: it starts with motion information in small receptive fields (e.g., motion of lines) and expands to higher-level receptive fields (e.g., overall direction of the object). While the motion pathway seems to be most relevant for biological motion understanding, Giese and Poggio's model proposes that the form pathway also plays an important role by using snapshots of poses to identify the types of action. Support for the role of the form pathway comes from studies that show that point-light display recognition is view dependent (e.g., changes in depth reduce recognition), which is a property of the ventral pathway (Bülthoff et al. 1998). Furthermore, some patients with damage to the motion pathway are still able to recognize biological motion, which suggests some role for the form pathway (McLeod et al. 1996). The motion pathway, however, is still the dominant system for recognizing point-light display motion, and fMRI work has found that the distinction between biological and nonbiological motion typically occurs in higher areas of the dorsal pathway (Decety and Grèzes 1989). In humans, it seems that multiple parallel pathways are involved in action understanding.

While biological motion is an important feature for identifying animate entities, it is also possible to identify these entities by virtue of their interaction with other objects. Evidence for this in infants comes from a study by Woodward (1998), who showed scenes where an arm grabbed one of two toys (Figure 16.5). During the test, the objects were switched and the arm grabbed either the same toy (old goal) or the new toy (new goal). Woodward found that five- to six-month-old infants were surprised when the arm went for the new goal over the old goal. This suggests that they think that the arm's motion is guided by a mental goal or preference for the old goal. Infants were surprised by an arm reaching for a new goal, but not by a mechanical claw, which suggests that they only attribute goals to the arm. Luo and Baillargeon (2005) have shown that a similar preference is present for a box when it moves in a self-propelled manner. They presented five-month-old infants a familiarization scene where a box moved toward a preferred object; later at test, they showed the infants the box moving toward the same object (old goal) or a novel object (new goal). They found a difference between the preference for the old goal and new goal was larger when the familiarization scene has two objects as opposed to one object. When there are two objects, the movement of the box shows that it prefers that object over the other one. When there is only one

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.

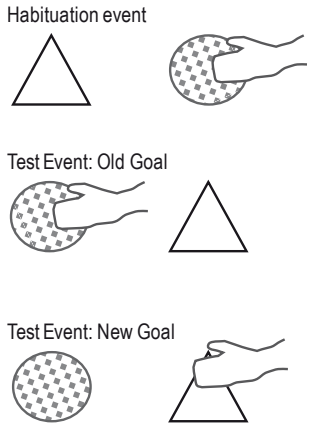

Figure 16.5 Depiction of goal recognition in infants, based on a study by Woodward (1998). Top: A child is shown an arm reaching out to grab an object (the circle). The position of the objects is then altered: the arm reaches for the same object (middle) versus a new object (bottom). When a human arm was shown, infants expected the arm to reach toward the circle. This expectation was not present when a mechanical arm was involved.

object, the movement toward that object does not show a strong preference, because there is only one option. Luo (2011) showed that even three-month-old infants demonstrate this preference, even when there was only one object at familiarization, as long as the box showed its preference for the object by moving toward the goal object in two different locations on the screen. In contrast to a self-propelled box, infants were surprised when an inert box reversed direction spontaneously, remained stationary when hit or pulled, or remained stable when released without support. One reason self-propelled motion is important is that the behavior of the box cannot be predicted based on physical constraints like gravity or inertia, and it is this unpredictability that triggers the assignment of mental states like goals.

The goal-directedness of human behavior was used to look at false belief behavior by Onishi and Baillargeon (2005). Here, 15-month-old children were first shown a familiarization event where a toy was hidden in a green box by a person in front of the infant. Then the scene was changed such that either the person's belief about the object in the green box would be false (e.g., the toy was moved secretly to the yellow box) or true (e.g., the toy was moved to the yellow box while the person watched). Later at test, the children observed a person reaching into the green box or yellow box. When that person saw the toy move into the yellow box, the children expected the person to reach for the yellow box and were surprised when that person reached into the green box. When the person did not see the toy move, the children were surprised when the person reached toward the yellow box, because they knew that the person had a false belief that the toy was in the green box. To show these differences,

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 infant cannot just record where the objects are located. They must also track where the person thinks these objects are located and how those beliefs, true or false, guide their reaching behavior. This is a nonlinguistic theory of mind task which shows that early in development, children can track the beliefs of others and use these beliefs to predict their behavior. Since ITL systems must interact with humans whose behaviors are driven by their beliefs (e.g., a teacher searches for a knife in a drawer, because that is where she thinks it is), it would be useful for these systems to have the ability to infer beliefs in the way that children seem to be able to do.

These abilities of infants to identify primitive mental states of others can support the understanding of more complex social motivations. For example, when six-month-old infants were shown a square trying to get up a hill, and it was helped up the hill by a triangle in one scene and hindered by a circle in another scene, they then preferred the triangle to the circle later, thus demonstrating that they understood that the triangle was helping the square reach its goal (Hamlin et al. 2007). By 18 months of age, toddlers and human-encultured chimpanzees are able to identify the ultimate goals of adult humans (e.g., putting books into a book case) and perform actions that help the adult to achieve these goals (e.g., opening the book case door) (Warneken and Tomasello 2006). Such helping behavior involves prediction of the goal, because the goal has not yet been achieved. In addition, the helper must identify objects that they can manipulate in the environment and predict whether these changes will help in achieving the goal. Since ITL systems are attempting to help humans in their tasks, these humanlike prediction abilities would enhance their interactions.

When a child comes to an ITL task, they have a model of the constraints that inanimate physical objects have in the world. In addition, they know that animate entities move in ways that reflect their goals and beliefs, and these biological entities can have multiple parts that work in concert to perform various actions. Furthermore, they can learn about behaviors in various contexts and use that knowledge to identify ways to help. While it is possible that children have a range of different modules which allow them to exhibit these capabilities, the range and flexibility of these abilities suggests that they are not separate isolated modules but rather part of a task-general system that integrates physical constraints, mental states, and learned regularities into a single system that attempts to predict behavior. Below, I propose that such a system will be useful for ITL.

A Developmentally Motivated World Prediction Model for ITL

How can ITL systems incorporate this rich database of knowledge that children seem to possess? In this section, I will suggest that incorporating a world prediction model can give ITL systems some of these abilities. To see how such a model might work, let us consider the following example: a robot learns

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.

from a human teacher how to cut a carrot. The robot has a carrot and a knife, and the human wants to show the robot how to cut the carrot by pantomiming a cutting action using her hand. For the robot to understand this action, it must be able to map the back-and-forth motion of the human hand in space with no carrot to its own hand with the knife. It would also need to know that the back-and-forth motion of the knife is the second component of a cutting sequence.

1. CONTACT: make contact between knife and carrot.
2. SAW-MOTION: move knife back and forth on carrot.
3. SPLIT: continue motion until carrot is in two pieces.

This task presents several challenges, from segmenting the event to understanding the pantomimed hand action. In pretend play studies, children from around the age of two years seem to understand pantomime actions of others and quickly generalize this knowledge to their own actions (Rakoczy et al. 2004; Rakoczy and Tomasello 2006; Rakoczy 2008). It would be useful for an ITL system to have a similar ability to understand pantomimed actions.

Human predictive knowledge is very detailed and context specific, so the world prediction model makes predictions at a fine time granularity. A model by Reynolds et al. (2007) does perceptual prediction to segment events; their model segmented sequences of routine actions encoded as point-light displays (as in Figure 16.4). They used a recurrent neural network that attempted to predict the next state of the points, using the previous state, and the error in prediction was used to update the model's weights so that it encoded the knowledge about the transitions between states. In addition to using error to learn the transitions, the model had an additional gating network that used points of large prediction error to identify event boundaries. To adapt this for ITL learning, we can assume that the system is attempting to use the present state of the world to predict the next state of the world (Figure 16.6). We will assume that the world state encodes static elements of the scene and the state of the robot's body that are derived from sensors and effectors (see Figure 1.1 in Mitchell et al., this volume). In addition, let us assume that the system tracks objects in the visual input, so that the world state is not a pixel-like representation but is instead based on static and motion properties of objects (e.g., shape, velocity, acceleration). Objects are any moveable element including inanimate artifacts (e.g., knife) as well as animate entities and their body parts (e.g., hand). The world state at time t is the input (bottom box, Figure 16.6); its activation is spread through internal layers with recurrent connections to help in the learning of sequential regularities; and the model generates a prediction for time t+1 at its output (second box from top, Figure 16.6). The actual world state at t+1 is the target (top box, Figure 16.6) and the difference between the target and the predicted world state is the error. The error passed back to the internal layers is used to learn representations that encode the internal parts of events. Importantly, the input state of the world at time t and t+1 is different from the knowledge of the world inside of the robot, which is encoded in the internal

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.

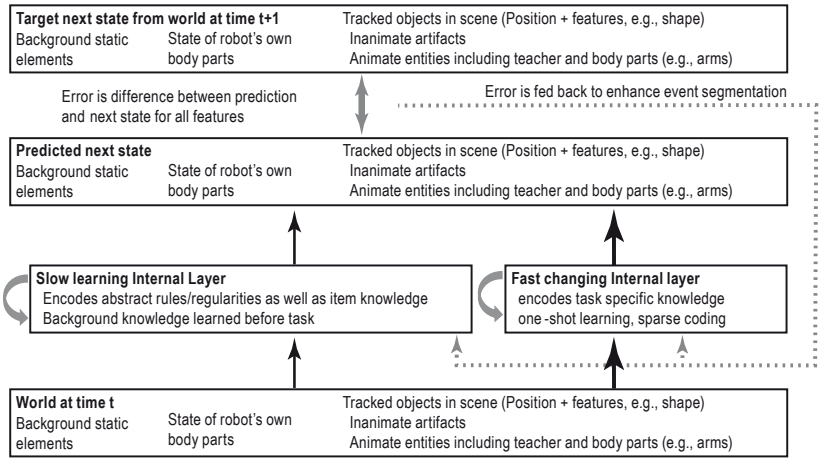

Figure 16.6 World-state prediction model.

layers. If we take the example of learning the carrot cutting sequence from a human demonstration, the model will be attempting to predict the state of the world multiple times a second. After it sees the back-and-forth movements of the knife (SAW-MOTION), the system should be able to predict that motion. But the system will not be able to predict that the carrot will split into two objects, and the large prediction error will be evidence for segmenting the SAW-MOTION component from the SPLIT component of the cutting sequence.

In ITL, there are aspects of the task that are specific to the particular situation (e.g., the knife is in the drawer) and aspects that reflect long-term knowledge about similar tasks (e.g., hands tend to reach for knives, but knives do not reach for hands). To model these different aspects of world knowledge, the world-state prediction model has two internal layers. The layer on the right side in Figure 16.6 is called the fast-changing internal layer: it has a high learning rate, which allows it to learn task-relevant temporary knowledge quickly. The layer on the left side of Figure 16.6 is called the slow-learning internal layer: it has a lower learning rate, which allows it to learn slowly regularities that are consistent across the whole of its previous input experiences. It has been argued that humans have a similar distinction in their brains between slow cortical learning and fast one-shot hippocampal learning (McClelland et al. 1995). We assume that the slow-learning layer develops its knowledge gradually based on many years of visual experience in the world. It would first learn to predict transitions based on physical constraints (Spelke et al. 1992). Given that the motion of self-moving objects is not fully predictable from physical constraints (Luo 2011), the prediction error that is generated will cause the model to learn separate internal representations for the actions of animate entities and these representations will eventually come to encode

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.

mental state features (e.g., goals, preferences, intentions), because these are useful for predicting the motion of these entities (Woodward 1998). After the slow-learning network has encoded background knowledge, then in an ITL task, the fast-learning network will only encode the aspects of the scene which are not predictable from the slow-learning network. For example, the slow-learning network has experience with hands reaching for objects, so it does not store the low-level motion of the hand. Because it has less experience with inanimate objects moving against another inanimate object (as when the knife is cutting the carrot), the prediction error generated causes it to be richly encoded in the fast-learning system. Thus there is a division of labor in how knowledge is distributed to the fast and slow subnetworks, as can be seen in connectionist models of language (Chang 2002; Janciauskas and Chang 2018).

Can this world-state prediction model be useful in understanding how an ITL robot would be able to understand a pantomimed cutting demonstration? Let us assume that the robot has grasped the carrot and placed the knife in contact with it, but is unsure how to cut the carrot so the human pantomimes the back-and-forth motion using her hand. The world prediction system cannot initially predict the back-and-forth hand motion based on its general world knowledge; hence the large prediction error that is created causes this motion sequence to be encoded in the fast-changing internal layer of the model. The world state is encoded in terms of the motion of objects independent of their features (e.g., shape, color), so the predicted sequence of states encoded by the fast-changing layer can be used to guide the motion of the robot's hand just by mapping the object pointer for the teacher's hand to the robot's motor control system (allowing it to exhibit childlike generalization, Rakoczy et al. 2004). Finally the large change of state that takes place when the carrot is split into two is used to segment the event, and this might cause the robot to stop and evaluate what to do next. Thus the world-state prediction model helps to explain how a robot with no knowledge about cutting events could learn the back-and-forth motion component of cutting sequences from interaction with a human teacher.

It is clear that a world-state prediction model would only be useful in an ITL robot if it was tightly integrated with other systems for planning and interaction (see Salvucci et al., this volume). It suggests, however, ways to apply previously learned background knowledge from non-ITL situations to support event segmentation and learning from visually taught events. Although the model is described within an error-based learning recurrent network, other algorithms could also be used. What is critical is that predictions are generated at each point in time, so that an error signal can be generated which identifies which aspects of the scene are unexpected. It is also important that the system learn its internal representations based on prediction mismatch, so that it has more extensive internal representations for less predictable entities like goal-driven agents. Furthermore, the prediction error is itself a signal that can be used to identify points where human feedback is needed or to segment events.

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 growing body of work in psychology argues that prediction is taking place all the time, particularly in language processing (Altmann and Kamide 1999). Young children seem to generate linguistic expectations automatically (Lew-Williams and Fernald 2007; Mani and Huettig 2012). Furthermore, changes that take place in the language representations of adult language users can be explained as prediction error-based learning (Chang et al. 2006; Dell and Chang 2014). If robots are doing a similar type of prediction about the nonlinguistic world and constantly updating their knowledge within a system like the world-state prediction system, then they would have rich moment-by-moment predictions and error signals, which could be used to learn new tasks. Thus, although the goal of ITL is not to model the development of human abilities in robots (i.e., developmental robotics) (Cangelosi et al. 2015), incorporating a module that is motivated by detailed prediction abilities of humans into these systems could enhance the ITL capabilities of artificial agents.

Conclusion

Humans engage in ITL based on multiple years of experience watching animate and inanimate entities interact. Children develop a model of how inanimate entities move based on physical constraints and how animate entities move based on inferences about their mental states. They can identify temporary goals in particular contexts and use long-term knowledge to understand the actions of others. Although ITL systems implement some of these abilities in separate modules, research on humans suggests that these abilities are the result of a system that is constantly involved in making predictions and adapting these predictions in response to experience (learning). Furthermore, prediction error may be a useful diagnostic signal for ITL systems. Given that humans expect these abilities when they teach and interact with other humans, it is likely that humans will prefer to interact with ITL systems that have a rich internal predictive model.

Acknowledgments

Franklin Chang is a member of the International Centre for Language and Communicative Development (LuCiD) at the University of Liverpool and the support of the Economic and Social Research Council [ES/L008955/1] is gratefully acknowledged.

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.