Coordinating Joint Action in a Real-Life Activity:

The Interplay of Explicit and Implicit Coordination

Chen Zheng

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2022

Abstract

Coordinating Joint Action in a Real-Life Joint Activity:

The Interplay of Explicit and Implicit Coordination

Chen Zheng

Humans engage in joint actions on a daily basis. Some of these joint actions are explicitly coordinated using, for example, speech and gesture, and the others are implicitly coordinated with the actions themselves. The first chapter of this dissertation reviews the use of speech, gesture, and intentional behavioral signals in explicit coordination of joint action and identifies three cognitive mechanisms that enable implicit coordination of joint action, namely, motor resonance, joint intentionality, and environmental and social affordance. The second chapter reports an empirical study exploring the employment of explicit and implicit coordination of joint action in a complex real-life joint activity, assembling a TV cart from its parts. We coded the content of the utterances and gestures that pairs of participants used throughout the assembly and the major and subordinate joint actions they performed. We then coded how each joint action was coordinated, that is, using speech, gestures, or action itself. The results showed speech and gesture served primarily to establish and sustain a shared mental model of the environmental affordances between the co-actors, which occurred primarily at the beginning of the task and as the participants began to attach two major parts. For both major and subordinate joint actions alike, the specifics of the joint actions such as the goal and division of labor was primarily coordinated implicitly. We argue that the shared mental model scaffolded the participants' implicit coordination of the actions. These findings provide evidence that action itself is a communicative device and part of the conversation between co-actors of a joint activity. They

also lend support to the argument that joint action cannot be fully understood on the individual level but must be interpreted as a collective of which each individual is a part.

Table of Contents

List of Figures iii

List of Tables v

Acknowledgments vi

Chapter 1: Explicit Devices and Implicit Mechanisms of Coordinating Joint Action 9

1.1 Explicit Devices of Coordinating Joint Action 11

1.1.1 Speech 11

1.1.2 Gesture 13

1.1.3 Signaling With Action 19

1.1.4 Discussion 21

1.2 Implicit Mechanisms of Coordinating Joint Action 21

1.2.1 Motor Resonance 22

1.2.2 Joint Intentionality 25

1.2.3 Environmental and Social Affordance 32

1.3 Conclusion 35

Chapter 2: Coordinating Joint Action in a Real-Life Joint Activity 37

2.1 Methods 41

2.1.1 Stimuli 41

2.1.2 Transcription and Coding 42

2.2 Results 50

2.2.1 Content of Speech 50

2.2.2 Type of Gesture 52

2.2.3 Establishment of Joint Mental Model 53

2.2.4 Coordination of Joint Action 57

3.3 Discussion 59

3.3.1 Explicit Coordination: Mental Model of the Environment 60

3.3.2 Implicit Coordination: Specifics of an Action 64

3.3.3 Sporadic Explicit Coordination of Action Specifics 65

3.3.4 Application in Human-Computer Interaction 66

Conclusion 72

References 73

Appendix A 84

Appendix B 88

Appendix C 90

Appendix D 91

List of Figures

Figure 1: A pair of participants in the middle of the task……. 40

Figure 2: Completed TV cart with its parts labeled, adapted from Heiser et al. (2004)……... 40

Figure 3: Coding scheme of content of speech……. 44

Figure 4: Average rate of each subcategory of utterances……..…...51

Figure 5: Element-establishing utterances…… 52

Figure 6: Rate of different gesture types……... 53

Figure 7: Occurrence of *part-and-structure* utterances and major actions……... 54

Figure 8: Occurrence of *action* utterances and major actions……54

Figure 9: Occurrence of part-and-structure utterances……. 56

Figure 10: Occurrence of action utterances……..…… 57

Figure 11: Coordination of major actions……. 58

Figure 12: Speech coordinating major actions.....58

Figure 13: Coordination of sub-actions…… 59

Figure 14: Speech coordinating sub-actions……..59

Figure 15: Ways the established elements were acknowledged……91

Figure 16: Scale of actions coordinated by action-element-establishing utterances…….92

Figure 17: Fate of action-element-establishing utterances……93

Figure 18: Rate of utterances across the four quarters of the task…….93

List of Tables

Table 1: Coding scheme of gesture……..……..45

Table 2: Coding scheme of action…….47

Table 3: Interrater reliability……..91

Acknowledgments

It is a blessing to research and be embraced by humanity. It takes lots of strenuous efforts to research, but the beauty of science and the sheer happiness of learning is incomparable.

During the past six years, I have had the privilege to investigate and learn about the interconnectedness of humans and to be mesmerized by the science and philosophy of *togetherness*. Everything started with my advisor, Professor Barbara Tversky, who gave me a goldmine of research topics, guided my every step forward, and gave me some of the most valuable suggestions in life. Her zest for life freed me up from the shackles of rigid theoretical thinking and led me to finding inspiration in the vast and vibrant wilderness of life. I also thank Dr. Erica Cartmill and Dr. Jacob Foster, who organized the fantastic Diverse Intelligences Summer Institute where I learned and set a bar for the kind of science I wanted to do and the scientist I wanted to be. I thank Professor James Corter, Dr. Bryan Keller, Professor Steven Feiner, and Dr. Christopher Baldassano, for all your valuable advice on structuring and writing up my dissertation. A special thanks to Dr. Baldassano, who witnessed and generously helped out on my pursuit of training myself into a cognitive neuroscientist. I also thank my friends Xu, Yuyan, Maqar Mehdi, Zhang, Jing, and Li, Lu, for helping with coding.

I thank my parents, Zheng, Jianhua, and Liu, Xiaoming, and my grandparents, Liu, Zuochun, and Song, Yanfang, for planting the seed of scientific pursuit in me and for your unconditional love and support. You are the source and origin of my courage and idealism. You are my harbor of ease. You are my reason and motivation for a better life.

I thank my teacher since childhood, Wang, Fuquan, my friends, Li, Lu, and Alexandra Reblando, and my decade-long friend since college, Si, Minqiang, for your unwavering support during the darkest times of this journey.

I thank my master of martial arts, Master Henry Moy Yee, his wife, Santy Moy, and everyone else in the Kung Fu family. Thank you for giving me a home half the earth away from home.

Lastly, I thank my dear friend and the other me in this world, Jonathan G. Bowen, soon-to-be Dr. Bowen, for introducing me to the splendid world of philosophy. I have greatly enjoyed every one of our conversations over science, philosophy, humanity, and life. Thank you, Jon, for always believing in me, for all the happiness we have shared, and for all your encouragement and company. Your friendship will always be held dearly in my heart, and I hope our paths cross again!

To my loving great-grandmother

Xu, Mingzhi

who was as young as I am now seventy-eight years ago

Chapter 1: Explicit Devices and Implicit Mechanisms of Coordinating Joint Action

Humans engage in joint activities on a daily basis, from simple actions such as avoiding collisions while walking and greeting a friend on the street to more complicated situations such as assembling a piece of furniture together and playing music in an orchestra.

Humans are naturally adapted and motivated to coordinate with one another in social life (Tomasello et al., 2005). Taking an ontogenetic perspective, around as early as 12 to 15 months, infants begin to proactively engage in coordinated activities with their mothers (Bakeman & Adamson, 1984). Our ability to organize our own behaviors and attention around others arises early in life. For example, in an experiment by Warneken and Tomasello (2007), infants at the age of 14 months were tested on whether they could (1) access an object placed inside of a transparent cylinder while a partner made this possible by holding the cylinder in place and (2) hold the cylinder in place so the partner could access the object. In both cases, success of the joint task depended primarily on the timing of the individual who held the cylinder in place. Children at this age could successfully perform the first task after viewing demonstrations but had difficulty in the second. Not surprisingly, however, their success rate in the second task improved significantly as they reached 24 months old (Warneken et al., 2006).

Humans are easily influenced by other individuals engaged in the same activity. In a study by Sebanz, Knoblich, and Prinz (2003), two participants sat next to each other while watching a finger on a screen pointing to the left (at one of the participants), center (between the participants) and right (at the other participant). The finger wore a ring colored red or green, and each participant was tasked with responding to one of the colors by pressing a button located on their own side. A spatial compatibility effect (Simon, 1990) was observed—the participants responded faster when

the finger pointed to their own side than when it pointed to their partner's side despite the finger's pointing direction being irrelevant to the task. These results suggested the participants formed a mental representation for not only their own but also their partner's part of the task.

A follow-up study (Sebanz, Knoblich, & Prinz, 2005) further suggested these mental representations were equivalent—the participants represented their partner's part the way they represented their own. In this study, participants responded to either one of the colors or one of the pointing directions, and they did so either jointly, sitting side by side and each responding to a different type of stimuli, or independently. If the participants represented the specifics of their partner's part equivalently to that of their own, their reaction time should be longer in the joint condition because the stimuli should have activated their mental representations of both sides' actions in the joint condition which would be in competition when they selected the correct response. The results confirmed this prediction—the participants responded slower when acting jointly and faster when acting alone.

Having a mental representation of each other's parts seems to help coordinate the co-actors' individual actions in a joint one. A joint action, by Sebanz, Bekkering, and Knoblich's (2006) definition, is a "social interaction whereby two or more individuals coordinate their actions in space and time to bring about a change in the environment" (p. 70). This highlights the point that it is coordination that ultimately makes individual actions a joint one (Clark, 1996).

In this chapter, we will examine how coordination is achieved to enable joint action. We will divide coordination into explicit and implicit coordination and discuss the devices of explicit coordination and the cognitive mechanisms permitting implicit coordination. By "explicit", we mean the signals sent by one co-actor express meanings that are meant to be directly interpretable by the other, and coordination, therefore, does not require making inferences about the joint action.

By contrast, implicit coordination does not involve intentional use of signals, rather, elements of a joint action are inferred from the collaborators' actions themselves, usually guided by a joint model between the co-actors concerning the common goal and the affordances of the objects involved and the world.

For the explicit devices, we will focus on the use of speech, gesture, and intentional actions in coordinating joint action, and for the implicit mechanisms, we will elaborate on the roles of motor resonance, shared intentionality, and environmental and social affordances. Eye gaze is presumably another useful coordinative device, but because the experimental stimuli we use in Chapter 2 do not allow accurate tracing of the co-actors' eye gaze, we will refrain from going into details about it and save the topic for future discussion. Additionally, for the sake of simplicity, we will not make specific distinctions between changes in the mind and changes in the environment elicited by a joint action. We will also limit our consideration of joint activity to the scale of dyadic interaction.

1.1 Explicit Devices of Coordinating Joint Action

1.1.1 Speech

1.1.1.1 Language as Joint Action and Coordinative Device

Conversation is a form of joint action itself (Clark, 1996). During a conversation, speakers monitor their own speech to make sure it is intelligible to their addressees, and the addressees provide evidence of their state of understanding so the speakers could adjust their subsequent utterances accordingly (Clark & Krych, 2004). Whenever an addressee provides sufficient evidence that they have understood the speaker well enough for current purposes, an alignment is achieved between the interlocutors, and the content of this alignment is "grounded" between them (Clark, 1996; Clark & Schaefer, 1989; Clark & Wilkes-Gibbs, 1986). On this account, language is

also a coordinative device whereby interlocutors align their understanding of the current situation (Clark, 1996; Garrod & Pickering, 2004, 2009).

Clark (1996) provided an illustrative example of language use during a customer-cashier transaction. The customer expressed their intention of making a purchase by saying “These two things over here.” then the cashier told him the total cost was “Twelve seventy-seven.” then the customer acknowledged this by repeating “Twelve seventy-seven.” and they proceeded with the payment process (p. 51). Here, by giving the total cost, the cashier demonstrated her understanding of the customer’s intention, and by repeating the total cost, the customer demonstrated his understanding of the cashier’s response. Both of them responded so their interlocutors would know they had reached an alignment over the current state of the transaction.

1.1.1.2 Coordinating Without Language

Despite its remarkable capability of aligning interlocutors’ mental representations, language may not be uniformly used in every part of a joint activity. Kobayashi et al., (2018) found participants’ use of both common nouns (e.g., box, car, etc.) and demonstrative pronouns (e.g., this, that, these, those) decreased as they repeated the same joint action over and over again. In more extreme cases, joint action could take place without language at all. For example, in a study by Richardson, Marsh, and Baron (2007, Experiment 4), pairs of participants were presented with planks of wood conveyed in an ascending, descending, or random order of length on a conveyor belt. The participants were instructed to lift the planks either by themselves, with the two hands grasping on both ends of the plank simultaneously, or jointly, each touching one end of the plank with one hand. Of interest was the finding that the participants switched between the two modes of lifting without any prior discussion.

Coordination without language was also observed in Clark and Krych's (2004) study where participants worked in pairs to assemble Lego models. One participant of each pair served as the director, verbally instructing the other (the builder) to build the models from loose Lego blocks. When the builder's actions were visually accessible to the director, grounding could be achieved silently, for example, a builder could demonstrate their understanding of the director's instruction by executing the very action that was instructed, or they could exhibit their lack of understanding or disagreement by not abiding by the instruction. The directors in this study appeared to adjust their instructions online according to the builders' manifest actions, which suggested language was not the only means of communication between interlocutors. We will come back to this point in the second half of this chapter.

1.1.2 Gesture

Similar to language, gesture is another tool of explicit communication. In this dissertation, we consider gesture as a manual action without an object. In a coordinative situation, speakers produce gestures mostly along with their speech (evidence provided in this dissertation), and listeners glean information from these gestures. Gesture thus seems to play a facilitative role in communication when it comes to coordinating joint action. Specifically, gesture fosters communication by directly enhancing the speakers' speech production and the listeners' speech comprehension (Driskell & Radtke, 2003; Hostetter, 2011).

In this part of the chapter, I will focus on the use of *illustrators*, spontaneous gestures that are part of the communication but do not belong to a codified system (as opposed to emblems such as thumbs-up, okay, and shush) (Ekman & Friesen, 1972; Goldin-Meadow & Wagner, 2005), in facilitating communication. To avoid possible confusion, I will use the word "gesture" for illustrators from now onwards.

1.1.2.1 Gesture Fosters Speech Production

Levelt (1989) divided speech production into two stages, conceptualization and formulation. Conceptualization refers to the stage at which the speaker determines the content of their utterances (what to express and what to omit in what order), and formulation refers to the process of grammatical and phonological encoding (Krauss, 1998). Gesture seems to facilitate both stages of speech production.

In an experiment by Melinger and Kita (2007, Experiment 1), participants were asked to describe pictures containing patterns of dots from memory. Some of these patterns were *deterministic* as the dots formed a single path along which the participants could follow, and the others were *non-deterministic* as the path had two branches and the participants had to decide which one to describe first. The branching added to the participants' cognitive load during the conceptualization phase, and, as the results showed, the non-deterministic pictures elicited more deictic (i.e., pointing, McNeill, 1992) gestures than the deterministic pictures, and the participants appeared to gesture the most at the branching points of the non-deterministic pictures. This suggested the gestures might have helped free up the speakers' processing resources during the conceptualization stage and thereby facilitated their speech production.

The theory by Krauss (1998) holds that gesture facilitates speakers' lexical retrieval, that is, the formulation stage of speech production. In a study by Frick-Horbury and Guttentag (1998), participants read definitions of uncommon words and tried to retrieve them from memory. Those with unrestricted hand gestures retrieved significantly more words than those who were restricted from producing gestures. In line with this result, Rauscher et al. (1996) found when speakers could not gesture, they produced a higher proportion of filled pauses (e.g., "um", "uh", "er") in the

middle of syntactic clauses than when they could gesture, which was a sign of lacking lexical access.

The facilitative benefits of gesture in both the conceptualization and the formation stages of speech production seem to suggest a shift of cognitive load from the verbal to the spatio-motor modality in working memory, which permits the speakers to allocate more cognitive resources to the verbal task (Goldin-Meadow, Nusbaum, et al., 2001; Wesp et al., 2001; Melinger & Kita, 2007). In cases like these, gestures are not produced to aid the listeners' understanding but as an inherent element of speaking itself.

1.1.2.2 Gesture and Speech Comprehension

Aside from gesture's facilitative effects on speech production, speakers also gesture (at least in part) intentionally to facilitate their listeners' comprehension. For example, speakers gesture more frequently to listeners who are naïve to the content of the discourse than listeners who are hearing the message for the second or third time, and they gesture more to listeners who have no visual access to the object being described than those who have a visual reference (Jacobs & Garnham, 2007). In real-life, it would be difficult and unnecessary to determine who the gestures are produced for, and the crux of the relationship between gesture and speech comprehension, therefore, is not how speakers make their gestures informative but whether viewing the speakers' gestures promotes the listeners' comprehension of speech. Evidence of the latter is mixed.

In their study, Krauss et al. (1995) asked participants to either watch or listen to video clips of a speaker describing an item. The item could be an abstract graphic design, a novel synthesized sound, or a sample of tea. The participants were then asked to select the item from a group of similar things. Contrary to what the authors had expected, the rate of correct selection did not differ between the watching and listening conditions; allowing listeners visual access to the speaker's

gestures did not enhance their efficiency of communication. What should be noted, however, was this study did not specify whether the speaker's gestures replicated their content of speech or provided additional information, and, if the latter, whether this information aligned or conflicted with the speaker's content of speech.

With this question clarified, Singer and Goldin-Meadow (2005) taught third- and fourth-grade children the concept of mathematical equivalence. The instruction was given verbally, accompanied by mismatching gestures, matching gestures or no gesture at all. The mismatching gestures aligned with the verbal instruction in content but had a different form (see also Goldin-Meadow, 2003), here providing an alternative strategy for solving mathematical equivalence problems, and the matching gestures replicated the speech they accompanied. The results showed children benefited from viewing the instructor's gestures only when the gestures mismatched the verbal instruction.

Mismatching gestures provide information uninvolved in their accompanied speech but relevant to its content. For instance, deictic gestures may serve to guide a co-actor's attention when referential words such as "this" and "that" are missing. Mismatching gestures seem especially useful in communicating dynamic information. For example, Kang & Tversky (2012) showed participants a step-by-step video explanation of the working process of a four-stroke engine. With identical verbal scripts, the gestures in each video either portrayed the movements of each part of the engine or pointed to their locations and delineated their shapes. In a subsequent test, participants who watched the movement gestures demonstrated a deeper understanding of the engine's operational mechanism as they depicted more specific actions of the system in the diagrams they made. Following the same vein, Hostetter's meta-analysis (2011) reported that

gestures are more communicative when they convey spatio-motor information and less so with abstract topics.

Although, as Hostetter (2011) pointed out, mismatching gestures seem to have larger communicative effects than matching gestures, it must be noted the effects of gesture could be modulated by a listener's own ability of speech comprehension. For example, when preschool and kindergarten children received instructions given in complex clauses, preschool children who had difficulty understanding the clauses benefited from viewing the instructor's matching gestures while kindergarten children who could easily understand the clauses did not (McNeil et al., 2000).

Thus, the facilitative effects of gesture on speech comprehension seem to be a function of the interplay between a listener's ability to understand the semantic content of the speech and the type of gesture they view; mismatching gestures seem to facilitate comprehension of discourses that are easily understandable, and matching gestures seem to facilitate comprehension of discourses that are hard to understand. In this regard, a listener must be able to understand both speech and gesture in order to achieve the best results of speech comprehension.

1.1.2.3 Commonality Between Understanding of Speech and Gesture

It turns out humans are adept at both speech and gesture understanding. Neurological studies have suggested speech and gesture understanding share a neural integration system on the semantic level. For example, both gestures and words that mismatch the semantic context of a sentence (e.g., He slips on the roof and *walks* to the other side.) were found to trigger an N400 effect (which is generally related to semantic mismatches) after stimulus onset, and this effect was similar in latency, amplitude, and topographical distribution for gesture and word mismatches (Özyürek et al., 2007). The N400 effect was also found triggered by words that mismatched the semantic context established by preceding gestures (e.g., Kelly et al., 2004). Moreover, gesture

and word mismatches appeared to activate the same brain areas in the left inferior frontal cortex, specifically, Broca's area and its adjacent cortex (BA 45/47) (Willems et al., 2007). Based on these results, it seems reasonable to argue speech and gesture understanding recruit the same neural resources for semantic processing.

In fact, Rizzolatti and Arbib (1998) found evidence suggesting a mirror system for gesture recognition including Broca's area in humans; in monkeys, neurons in the mirror system discharge when a monkey manipulate objects and when it observes similar actions. Rizzolatti and Arbib argued that the mirror system in humans provided a bridge between action and communication.

Tomasello (2010) further argued that gesture might be the evolutionary origin of human language, which is evident in the flexibility of gesture use in great apes that meet different social ends across different social situations. For example, great apes are sensitive to the recipients' attention while producing gestures; they gesture almost exclusively when the recipients' attention is already oriented at them, otherwise they would walk around to meet the recipients' attention (Call & Tomasello, 2007). Great apes also monitor the recipient's reaction to their gestures, and they would repeat the same gesture or use another gesture if the desired reaction were not elicited. Some human-raised apes even learn new gestures specifically for interacting with humans, for instance, they would solicit help from humans by pointing imperatively to specific targets. However, great apes do not gesture cooperatively other than making imperative requests—they do not point to share attention or to inform another of something helpful, and when humans do this for them, they cannot understand the declarative or informative intentions behind. By contrast, human infants start sharing attention and informing others through gesture early on in ontogeny, and, as Tomasello (2010) argued, the skills and motivations of such cooperative communication constitute the infrastructure of human communication.

To summarize, humans are proficient gesturers and interpreters of gestures. In a joint activity, gesture facilitates communication by enhancing the speakers' speech production and fostering the listeners' speech comprehension, although how exactly the facilitative effects take place is determined by the listener's own language skills as well as whether the gesture overlaps with the content of its accompanying speech.

1.1.3 Signaling With Action

Both speech and gesture are signals—deliberate actions by which one individual conveys meaning for another to identify and understand (Clark, 1996; Clark & Krych, 2004). In this sense, the goal of signaling is to influence others through communication. Different from pure signals such as speech and gesture though, some actions are pragmatic as well as communicative—they fulfill a pragmatic goal in parallel with their communicative purposes (Pezzulo & Dindo, 2011; Pezzulo et al., 2019). For example, members of a string quartet may exaggerate their body movements to coordinate the tempo with one another, but, ultimately and pragmatically, their bodies are maneuvered to create the music.

1.1.3.1 Signaling With Emphasized Kinematic Cues

As in the example of the string quartet, signaling with pragmatic actions usually takes advantage of emphasized kinematic features of the actions. In an experimental setting (Sacheli et al., 2013), pairs of participants were asked to grasp a bottle-shaped object in synchrony. One of the two participants in each pair received auditory information about where to grasp the object (on the neck or the body) and, thus, became the leader, and the other participant, the follower, was instructed to complement the leader's actions (to grasp the body of the object when the leader grasped its neck and vice versa). Analyses of the kinematics of the participants' actions showed leaders sent kinematic signals to their partners to reduce asynchrony within the pair. For example,

leaders would reduce their grip apertures (i.e., the distance between the index finger and the thumb) when they reached for the neck of the bottle (see also Vesper & Richardson, 2014), and they would exaggerate their trajectories of reaching the bottle, going higher while aiming at the neck and lower while aiming at the body.

Not surprisingly, kinematic signals were sent as the result of a decision-making process. Leaders would monitor the pairs' performance history and adjust their signaling strategy online; kinematic signals were sent to proactively facilitate the followers' anticipation of the leaders' actions, and the leaders would stop sending signals when they inferred their partners could reliably predict their subsequent actions (Candidi et al., 2015). Similarly, when human participants collaborated with a computer model in a joint task (Pezzulo & Dindi, 2011), they weighed the cost of signaling (e.g., slower speed in action, uncomfortable hand movement, etc.) against its collaborative benefits (e.g., more accurate prediction of the human's actions by the computer model) and the degree to which these benefits could be successfully elicited by signaling.

1.1.3.2 Signaling Through Different Modalities

Indeed, coordination of joint action is a process of information sharing that can take multiple forms and be achieved through different modalities (Clark, 1996). As much as pianists rely on auditory feedback to coordinate a duet, they produce ostensive visual cues (e.g., higher finger movements above the key and more synchronized head movements) in compensation when auditory feedback is inhibited (Goebl & Palmer, 2009). Haptic signals could also be used for coordinating joint action. For example, in a joint manual pulling task (van der Wel et al., 2011), participants moved a pendulum-like pole back and forth by pulling a cord collaboratively on its opposite sides, thus creating a haptic channel of communication in addition to vision. Over time,

the pattern of the pole's movement resembled that normally found when individual participants manipulated the pole bimanually.

1.1.4 Discussion

In this section, we reviewed the use of three explicit devices in coordinating joint action, namely, speech, gesture, and signaling with pragmatic action. Speech is a form of joint action itself which could also serve to align mental representations between co-actors, and the use of speech typically decreases with the progress of a joint activity. Gestures are most often found accompanying speech, and they foster co-actors' communication by enhancing the speakers' speech production and the listeners' speech comprehension. Communicative information could also be embedded in the specifics of the co-actors' pragmatic actions, which could take place not only visually but also through other sensory modalities.

1.2 Implicit Mechanisms of Coordinating Joint Action

Unlike signals, some actions are not intended for conveying meanings but are happenstances when co-actors act in a specific context—what Clark and Krych (2004) called *symptoms*—nevertheless, collaborators can take advantage of these symptoms and coordinate their joint actions accordingly (Clark & Krych, 2004; Shintel & Keysar, 2009). We will call such symptom-based coordination "implicit coordination" as it does not implicate using any intentional signal.

Implicit coordination is not a rare phenomenon in our daily lives. As we navigate a crowded street, pedestrians do not signal one another to avoid collisions (although they sometimes do); at a grocery store, it does not take signaling to coordinate the lines unless their shapes are unclear. On a more social level, there are social etiquettes that people assume during conversations, in the

classroom, on the dining table; roles imply status and behaviors (e.g., first violins lead string quartets); rituals and conventions prescribe actions (e.g., weddings and other ceremonies).

In this section, we will identify three cognitive mechanisms that potentially underpin implicit coordination of joint action, and we will focus on the coordination of joint actions that occur in face-to-face instantaneous joint activities such as assembling a piece of furniture and preparing a meal.

1.2.1 Motor Resonance

The first mechanism, motor resonance, refers to "the activation of the motor system during action observation" (Uithol et al., 2011, p. 2). The ideomotor theory (Prinz, 1997) postulates that observing an actor executing an action activates the observer's own motor representations of the observed action which include both the motoric movements and the goals they attain.

The initial neurological evidence of motor resonance comes from mirror neurons originally discovered in the ventral premotor cortex of macaque monkeys (di Pellegrino et al., 1992; Gallese et al, 1996; Rizzolatti, Fadiga, et al., 1996). These neurons fire both when a monkey *performs* an action and when it *observes* the action being performed by another monkey or a human. In humans, fMRI studies have identified a parieto-frontal mirror circuit (for reference, see Rizzolatti & Craighero, 2004) in areas analogous to those in macaque monkeys. Interestingly, mirror neurons in macaque monkeys respond only to target-oriented actions (Rizzolatti & Sinigaglia, 2010), while in humans, the mirror circuit responds to movements both with and without a target (e.g., Decety et al. 1997; Buccino et al. 2001; Calvo-Merino et al., 2005).

1.2.1.1 Priming Observed Actions

An implication of the ideomotor theory is that perceiving the execution of an action would prime the observer for the same action; the initiation of an action would become easier and more

probable after observation (Wilson & Knoblich, 2005). This is supported by an experiment (Brass et al., 2001) in which participants, upon seeing a trigger on the screen, either lifted or tapped the index finger as per instruction. The trigger was a finger going randomly up or down, and the participants were instructed to perform a required finger movement irrespective of the direction of the observed finger movement. The results showed the participants responded faster when the direction of the observed movement was compatible with that of the required finger movement, a pattern remaining the same regardless of whether the palm faced up or down. The same priming effect was also observed in target-oriented actions, for example, participants grasped an object faster when they watched the experimenter grasp the same object first than when they did not (Castiello et al., 2002; Edwards et al., 2003).

The tendency to perform an observed action could be modulated by higher-order factors such as expectation, intention, and goal (Roepstorff & Frith, 2004). For example, Lakin and Chartrand (2003) asked participants to watch a confederate perform clerical tasks such as answering the phone and typing at the computer. Those who were told in advance to cooperate with the confederate at a later time mimicked the confederate's face-touching actions more than those who did not receive the instruction and who, therefore, did not have a cooperative intention. The increased mimicking of face-touching given a cooperative intention could be in the service of creating a cooperative foundation for further communication and/or joint action.

1.2.1.2 Priming Complementary Actions

In a joint activity, participants could be primed by observing each other's actions (Knoblich & Sebanz, 2008), which would foster their coordination especially if the participants worked in parallel and could imitate each other—for example, when lifting a heavy couch on different ends and dancing mirrored movements in a partner dance. However, not all joint actions are imitative,

rather, a lot of joint actions require complementary coordination where each co-actor is supposed to undertake a different part. Even those imitative joint actions may have a complementary aspect, for instance, individuals jointly lifting a heavy object are haptically coupled, and so are dance partners dancing together.

There is abundant evidence that complementary actions could also be primed through observation—provided that the observer had a complementary intention. In a study by van Schie et al. (2008), participants were asked to either imitate or complement a virtual co-actor's way of grasping a manipulandum. Under both conditions, 40% of the trials showed a colored cue signaling the participants to grasp the manipulandum in a pre-instructed way irrespective of the condition rule. For these trials, participants in the imitative condition responded faster when the observed action was compatible with the pre-instructed action, thus replicating findings mentioned above, but their response in the complementary condition turned out to be slower when the two actions were compatible and faster when they were incompatible. The reversed pattern of response suggested the intention to complement an action could override the tendency to imitate an observed action and foster the performance of the complementary action instead.

Results from an fMRI study (Newman-Norlund et al., 2007) further supported this possibility. Using the same paradigm as in van Schie et al. (2008), the authors found although the bilateral inferior frontal gyri (IFG) were activated in both conditions, the right IFG was more activated in the complementary context than in the imitative context. Considering the IFG's sensitivity to goals, the authors reasoned that the left IFG might have been activated by a shared overarching goal of both conditions—to generate an appropriate response to the signaling cue—and the right IFG might be sensitive to the more nuanced distinctions between them, for example,

to distinguish one's own subordinate goal from the partner's subordinate goal (the complementary condition) or not (the imitative condition).

Taken together, motor resonance offers a cognitive substrate that associates one's own actions with those of the others'. Observing another individual perform an action can facilitate one's own performance of the same action, but having a complementary intention may override this imitative tendency and, instead, foster the initiation of the complementary action.

The modulating effect of a complementary intention highlights the role of mentalizing (i.e., theory of mind) in joint action, that is, the ability to read the other individuals' desires, intentions, beliefs, and so on (Frith & Frith, 2008). The following part of this section will discuss the role of joint intentionality (of which theory of mind is a part) in coordinating joint action.

1.2.2 Joint Intentionality

The joint-ness of a cooperative joint activity consists in both aligned (e.g., audience clapping and soldiers marching in unison) and complementary actions (e.g., passing and taking, leading and following), and a joint goal seems critical for coordinating joint action when both parties intend to cooperate (see also Frith & Frith, 2008) as it regulates co-actors' individual actions in both domains.

In this chapter, I would like to consider the distinction between *goal* and *intention* proposed by Tomasello et al. (2005): Intention is a series of planned actions that an agent adopts in their pursuit of a goal. Then, having a joint intention implies (1) each co-actor plans their individual actions with respect to the other's intentionality, and (2) the co-actors plan their joint actions to achieve a joint goal.

1.2.2.1 Inferring and Responding to a Partner's Individual Intentionality

Humans code their partners' actions in terms of both what they are (e.g., a grasp) and why they are performed (e.g., grasping to remove) (Uithol et al., 2011), and there has been a growing body of evidence supporting this claim.

On the neurological level, mirror neurons in the inferior parietal lobule (IPL) of macaque monkeys exhibit different discharging patterns at the sight of actions serving different goals such as grasping to eat vs. grasping to place (Fogassi et al., 2005), and they discharge before the goal of the action can be visually determined. Similarly, a neuroimaging study with human participants (Koul et al., 2018) found different spatial patterns of the observers' brain activity in the inferior parietal lobule (IPL), superior parietal lobule (SPL), inferior frontal gyrus (IFG), and middle frontal gyrus (MFG) when they viewed the reaching part of a grasping action with different intentions (to drink or to pour from a water bottle).

On the behavioral level, the human eyes could distinguish between different intentions of an action on the sole basis of kinematic information. For example, using the same reach-and-grasp materials as in Koul et al. (2018), Cavallo et al. (2016) found participants could discern the intentions of a hand (grasping to drink or grasping to pour) from kinematic cues such as its wrist height, forward movement of the dorsum plane, and the horizontal wrist trajectory during the reaching phase. Social intentions (e.g., a co-actor being cooperative vs. competitive) were also discernible from early-stage reach-and-grasp movements, both when the participants viewed regular overall recordings of the hand movements (Sartori et al., 2011) and when they watched barebones point-light displays (Johansson, 1973; Manera et al., 2011).

In a joint activity, co-actors must know how to respond to their partner's intentions to achieve successful coordination (Becchio et al., 2010; Georgiou et al., 2007). In some cases, the

tendency to respond to a partner's intentions could be so strong that it could alter an actor's original plan of action. For example, in Becchio et al.'s (2008) experiment, participants were instructed to grasp an object and put it in a container. In some trials, a confederate would unexpectedly stretch an arm out and unfold their hand at the beginning of the participant's action as if to ask for the object from them. Under such "requests", the participants' arm trajectories veered significantly toward the confederate early on in the reaching phase. The trajectories also deviated toward the confederate after the object had been picked up; some participants even ignored the instruction and placed the object in the agent's hand.

In a study that manipulated social intentions (Sartori et al., 2009), participants were asked to either cooperate or compete with a partner in a tower-building task. In the cooperative condition, co-actors were supposed to put one block on top of another in succession, and, in the competitive condition, they were asked to compete to put the first block down. An experimental confederate served as one of the co-actors in each pair, and the confederate assumed an attitude that was either congruent or incongruent to the participant's expectation (i.e., to cooperate when the participant expected to cooperate or to compete when the participant expected to cooperate). When the confederate adopted an incongruent attitude, the participants' kinematic patterns were affected—cooperative participants appeared more competitive, showing larger maximum grip apertures, faster movement, and lower wrist trajectories, and competitive participants appeared more cooperative. In other words, they became more like the confederate when an incongruent attitude was detected.

The sensitivity to another's intentions and the tendency to respond to another's intentions seem to have their evolutionary origins. In a study by Warnekenet al. (2007), chimpanzees were tested whether they would lend help to an unfamiliar human or conspecific when they tried to

reach an out-of-reach target object but failed. The results showed chimpanzees helped both humans and conspecifics repeatedly regardless of any expectation of reward and even when effort was required (e.g., climbing up into a raceway to fetch the target), which was much the same way human infants behaved when they were tested with unfamiliar humans. Such results indicate not only humans but also chimpanzees are motivated to help another with their unachieved goal, which can only be possible whey they understand the others' intentions including how they themselves might be a part of them.

However, even with the sensitivity to others' intentions and prosocial motivations, is it necessary for co-actors to code each other's intentionality in a joint activity? After all, studies have shown coordination could be achieved through simple dynamical coupling between individuals; for example, people tend to synchronize in phase or out of phase when they swing a leg (Schmidt et al., 1990) or a pendulum (Schmidt & O'Brian, 1997) or rock in a rocking chair together (Richardson, Marsh, Isenhower, et al., 2007), provided they are visually available to each other.

I would like to argue the necessity to code the partner's intentionality should be analyzed against the nature of the joint activity. Swinging legs and pendulums and rocking in rocking chairs are all repetitive rhythmic actions, and the participants in these studies were all asked to perform the actions at their own comfortable pace (Schmidt et al., 1990; Schmidt & O'Brian, 1997; Richardson, Marsh, Isenhower, et al., 2007), which means they never had a common goal, and their synchronization was not so much a means toward a common goal than the sheer consequence of performing the actions. In contrast, coordination in a goal-oriented situation is the approach to achieving the joint goal and part of the participants' shared intentionality, and coding the other's intentions is, therefore, essential to the achievement of coordination.

Then the question is, why do people share intentionality in a joint activity?

1.2.2.2 Sharing Intentionality

1.2.2.2.1 Common Goal and Joint Planning

To answer this question, we first need to consider what shared intentionality means. According to Tomasello et al., (2005), shared intentionality refers to when co-actors engaged in collaborative interactions have a shared goal and a shared plan to pursue this goal.

In terms of *performing* a joint action, shared intentionality might indeed be unnecessary. Young children (Tollefsen, 2005) and nonhuman primates (cf. Horschler et al., 2020) who are usually regarded as limited in their mentalizing abilities—and, therefore, do not share intentions (Tomasello et al., 2005)—nevertheless engage in coordinated activities in one way or another. For example, infants as young as 12- to 24-month-old could give and take an object and roll a ball back and forth with an adult (Hay, 1979; Hay & Murray, 1982); chimpanzees cooperate in food retrieval (Melis et al., 2006) and offer help when a conspecific is having a problem (Yamamoto et al., 2012). Coordination seems possible as long as the co-actors understand the social affordances (action possibilities, Marsh et al., 2009) they provide for each other (see Chaminade et al., 2012, for fMRI results in support of this notion).

Shared intentionality subsumes joint planning and coordinating complementary roles and actions. Once a common goal has been established, participants of a joint action can start working backwards to find a plan that leads toward the goal, and they usually wind up adopting complementary roles (Clark, 1996; Knoblich & Sebanz, 2008). As mentioned above, a joint plan could be coordinated both explicitly (with speech, gesture, bodily signals, etc.) and implicitly. If implicitly, the common goal serves as a reference against which the co-actors could infer their partners' intentions and select appropriate actions in response (Pezzulu & Dindo, 2011). Of course,

there may be a hierarchy of goals in any joint activity, and the same logic should apply to each level of them.

Knoblich and Jordan (2003) conducted an experiment illustrating the process of implicit planning while participants shared a common goal. In this experiment, two participants jointly controlled a tracker that followed a target's movement along a horizontal line. The participants were each in charge of an accelerating button or a decelerating button, and the tracker would stop moving if no button was pressed. The target moved with constant velocity and would turn abruptly around after reaching either border, so the best strategy for the tracker was to execute an anticipatory brake—to decelerate before the target reached the border—which required pressing the decelerating button more than the accelerating button. The results showed the participants performed anticipatory brakes more often when they received auditory feedback of each other's pressing behaviors than when they did not. One possibility was the participants could detect each other's intentions from the auditory feedback and adjust their own pressing accordingly. Consistent with this view, receiving the auditory feedback did lead to a lower rate of compensatory key pressing which served to offset a partner's unwanted efforts. In short, for the participants in this experiment, their shared goal was to brake the tracker in anticipation of the target's turn which was achieved by predicting each other's intentions and selecting the appropriate complementary actions on their own ends.

1.2.2.2.2 The *You* and the *I*

Coordinating complementary actions requires keeping the *you* and *I* apart (Tomasello et al. 2005)—participants must distinguish their own part from their partner's part to achieve coordination (Knoblich & Sebanz, 2008). Specifically, one must suppress their tendency to perform an action when it is their partner's turn.

In favor of this view, Sebanz et al. (2006) found neural evidence of action inhibition in performing join action. In their study, participants were each assigned a color and asked to press a button upon seeing a colored cue on the screen. EEG data showed a larger P300 amplitude (at the frontal and central electrodes) triggered when the participants worked in pairs than when they worked alone. The authors argued perceiving a signal linked to the other's response activated the observer's own motor representation of executing the same action (motor resonance), and the increased P300 amplitude reflected the additional executive control demanded for inhibiting the tendency to actually perform the action.

Despite the practical necessity to keep one's own and the partner's parts apart, co-actors seem to monitor their partner's performance (Bekkering et al., 2009) the way they monitor their own (Clark, 1996; Prinz, 1997). On the behavioral level, people slow down not only when they themselves make a mistake but also when a co-actor makes a mistake (Schuch & Tipper, 2007). On the neural level, activity in the posterior medial frontal cortex (pMFC), an area concerned with internal monitoring of one's own action (Amodio & Frith, 2006), is increased when one observes a partner making an error (Bekkering et al., 2009). Thus, the monitoring of one's own and a partner's action seems to operate on the same cognitive and neural mechanisms, which, supposedly, is beneficial for co-actors engaged in a joint task in practice because the partner's actions would then provide an efficient confirmation or disconfirmation of one's own representation of the task including the shared goals and plans (Pezzulu & Dindo, 2011; Sebanz & Knoblich, 2009).

Therefore, the distinction and assimilation of the *you* and the *I* seem to co-exist in coordination; distinct roles and parts serve to ensure the performance of a joint action, while a coordinated *us* consists in the shared goal that guides the co-actors' individual actions.

In summary, although coordination may be achieved without shared intentionality between co-actors, joint planning cannot. Implicit planning is one type of joint planning when co-actors share a common goal and have perceptual access to the results of their joint actions. Joint goals and joint plans guide the co-actors' interaction with the environment as well as between the co-actors themselves.

1.2.3 Environmental and Social Affordance

As is clear from the above sections, during a joint action, the environment and the partner provide a co-actor with action opportunities (and constraints given their action repertoire) which, from an ecological perspective (Gibson, 1979/2014), are called their "affordances".

1.2.3.1 Automatic Perception of Environmental Affordance

Humans' perception of environmental affordances seems fast and involuntary. In an fMRI study (Grèzes et al., 2003), participants performed either a power grip (with the palm) or a precision grip (with the thumb and index finger) to indicate their categorization of an object (natural vs. man-made). This motor categorization was crossed by the size of the object—larger objects afforded power grip and smaller objects afforded precision grip—and the participants responded faster when an object's categorization was congruent with its affordance and slower when its categorization was incongruent with its affordance. Additionally, larger differences in reaction time between congruent and incongruent trials correlated with enhanced brain activity in areas involved in action selection (e.g., the anterior parietal, dorsal premotor and inferior frontal cortices). The authors argued both longer reaction time and greater brain activation indicated greater competition between the action opportunities specified by the category and affordance of an object, and the latter seemed to automatically activate the participants' motor representations of possible ways of interaction.

In another experiment investigating the effects of viewing a context (Iacoboni et al., 2005), participants viewed video clips of (1) objects (a teapot, a mug, cookies, a milk jar, etc.) arranged as before or after having tea (the Context condition), (2) a hand grasping the mug without a context (the Action condition), and (3) a hand grasping the mug within the contexts of the Context condition (the Intention condition). Interestingly, the Context condition, despite showing objects only and not any action, activated the inferior frontal areas (where mirror neurons were located) related to grasping, and the Intention condition, compared to the Action condition, elicited stronger activation in the caudal inferior frontal gyrus in the right hemisphere. These results suggested the context activated the participants' mental representations of potential ways of interaction between the hand and the objects in the environment.

1.2.3.2 Environmental Affordance and Joint Attention

Motor representations activated by environmental information is facilitatory to joint-action coordination in at least two respects. First, given that activation of one's own motor representations facilitates understanding of another's actions, motor representations activated by environmental affordances should facilitate a co-actor's understanding of their partner's actions in the same environment. Second, when multiple actions are afforded by the environment, these mental representations may code the most likely sequence of actions in this context (see also Jacob, 2008); contextual information might constrain which actions could be sequentially related to an observed action and thereby facilitate prediction of future actions.

Thus, for environmental affordances to facilitate co-actors' understanding of each other's actions in a joint activity, the best strategy is to align their perception of the environment. Aligning co-actors' perception of the environment requires joint attention (when both attend to the same aspects of the environment), and, for a self-organized pair, this most likely means they know they

share these perceptual experiences (Knoblich & Sebanz, 2008; Tomasello & Carpenter, 2007). Joint attention could be achieved by explicitly guiding another's attention through, for example, deictic gestures and vocal signals. Bodily information such as eye gaze (Flom et al., 2006) and body orientation (Jellema et al., 2000) could also be used to implicitly infer what another could and could not see.

1.2.3.3 Social Affordance

It should be noted that co-actors' perceptual experience during a joint activity is not confined to the physical environment; their partners are part of the environment in which they are involved. Partners in a joint activity provide social affordances in terms of what the pair could do relative to what each individual could do (Knoblich & Sebanz, 2008). The idea of social affordance was illustrated in the experiment by Richardson et al. (2007) mentioned earlier. In this experiment, participants lifted planks of wood presented in ascending, descending, and random orders of length from a conveyor belt. Although participants were present in pairs in each trial, they were free to lift the planks either individually with two arms spanning over the plank or jointly with each participant touching one end of it. The point at which the participants switched between solo and joint lifting was a function of their combined arm span instead of either participant's individual arm span, which suggested they took each other's capacity into consideration while executing the joint lifts.

Social affordance includes not only bodily affordances such as arm span but also an individual's agency in the sense of "behavior affords behavior" (Gibson, 1979/2014, p.127). Taking the simple interaction with a cup for example, a *passing* action affords *taking*, and a *displaying* action affords *viewing*; whether an actor takes or views the cup depends on whether their partner passes or displays it to them. Co-actors in a joint action appear to anticipate the action

opportunities afforded by their partner's actions. For example, Kourtis et al. (2013) found during a give-and-take task, the designated "takers" developed a slow negative brain potential upon the sight of a common cue which signaled the co-actors to get ready, and the potential peaked as the "givers" executed their giving actions. As Gangopadhyay and Schibach (2012) proposed, to perceive an individual's intention is to perceive a type of social interaction it affords.

Aside from direct interactions, a co-actor might mediate another's interaction with the environment by changing the environmental affordances. For example, *A* could make an object available to *B* by passing it over; *A* could make an object unavailable to *B* by taking it away; *A* could alter features of the object so *B* could no longer interact with it in certain ways but may have new opportunities. One individual's interaction with the environment causing changes that influence another's subsequent interaction with the environment is the stigmergy (a mechanism of self-organization observable in insects and animals when a trace left in an environment stimulates subsequent actions by the same or different agent, Heylighen, 2016) of human joint activity.

1.3 Conclusion

This chapter reviewed the cognitive and social mechanisms through which joint action could be coordinated explicitly and implicitly. In the first section, we discussed the use of speech, gesture, and action as explicit coordinative devices, and in the second session, we identified three cognitive mechanisms underpinning implicit coordination, namely, motor resonance, joint intentionality, and affordances.

Many questions remain unanswered. For example, what aspects of a joint activity are coordinated explicitly and implicitly? What factors determine how they are coordinated? How does explicit and implicit coordination work in concert? The next two chapters of this dissertation

aspire to shed light on these questions with behavioral data collected from a real-life joint activity, assembling a TV cart from its parts.

Chapter 2: Coordinating Joint Action in a Real-Life Joint Activity

As mentioned in the first chapter, the human life abounds with joint action. We coordinate with friends and family to prepare a meal, to clean the house, to dance to the music; we coordinate with colleagues in professional practices such as organizing a meeting and performing a surgery; we coordinate with computers and AI systems to translate languages, complete transactions, and, in the near future, travel in the streets.

The coordinative mechanisms of joint action have received abundant attention in recent decades. In his book, Clark (1996) proposed the idea of *common ground* where information is coordinated between co-actors.

A common ground could be efficiently achieved through the use of language. During a conversation, speakers monitor their speech to deliver intelligible utterances to their addressees who would, in turn, provide evidence of their state of understanding. For example, one co-actor could suggest a division of labor between the two of them or set a goal for the pair, and the other could confirm their agreement by saying "Okay." or request more information by asking a question. Interlocutors eventually achieve an alignment of understanding through loops of feedback (Clark & Krych, 2004); whenever there is sufficient evidence that the co-actors have understood each other well enough for current purposes, an aligned common ground is achieved between them.

Gesture is another means of coordination, usually as an accompaniment to speech. Gesture facilitates both the speakers' speech production (e.g., Melinger & Kita, 2007; Rauscher, Krauss, & Chen, 1996) and the addressees' comprehension (e.g., Singer & Goldin-Meadow, 2005). Some of these gestures are representational, standing for their referents (e.g., an action, entity, space, etc.) through pantomiming actions, isomorphous delineation, or reference of an empty space (Kita,

2000). Deictic gestures often serve to direct an addressee's attention to "an object, place, or event" (Clark, 2005, p. 509), which is achieved through pointing, touching, or manipulating an object with the gesturer's finger, hand, or other body parts.

Both speech and gesture are signals, deliberate actions by which one conveys meaning for another to identify and understand (Clark, 1996; Clark & Krych, 2004). Other forms of signals commonly seen in coordinating joint action include ostensive visual cues (e.g., exaggerated bodily movements, Goebl & Palmer, 2009) and haptic signals (e.g., van der Wel et al., 2011).

Despite their variety and prevalence, signals might not be necessary for coordinating joint action. For example, in a study by Richardson, Marsh, and Baron (2007), pairs of participants were presented with planks of wood conveyed in ascending, descending, and random orders of length on a conveyor belt. The participants were instructed to lift the planks either by themselves, with the arm spanning over the plank, or jointly, with each participant grasping one end of the plank, and they switched between the two modes of lifting without any prior discussion. In a study of collective musical improvisation, shared intention (e.g., to end a piece) emerged spontaneously during the course of improvisation and guided the coordination of musical dynamics (e.g., loudness) and harmony (Goupil et al., 2021). In real life, supermarket clerks know customers intend to purchase the items they place on the checkout counter (see Clark, 2005), volleyball players rotate after every side-out, audience line up to enter theaters—all coordinated implicitly without explicit communication between the co-actors.

In these cases, actions are not intended for conveying meaning but are happenstances when co-actors act in the same context in the service of a common goal—to lift the planks, to play music, to pay for groceries, etc. Nevertheless, co-actors could take advantage of their actions and

coordinate with each other accordingly. In this paper, we refer to this type of coordination of joint action as *implicit coordination* as it does not implicate any intentional signal.

Many studies have illustrated the phenomenon of implicit coordination of open-ended joint action. As mentioned above, Goupil et al. (2021) observed spontaneous emergence of shared intention during collective musical improvisation. In Roberts and Goldstone's (2011) study, participants each submitted a guessed number so the sum of the group matched a target value. No communication was allowed, but the participants assumed complementary roles over repeated rounds of play; some chose to stay inactive, adjusting their guesses on a minimal level, while the others continued to make bigger adjustments. In both real-life musical improvisation and laboratory-based number-guessing, co-actors must practice with each other until a convention is established from their actions.

To our best knowledge, no research has systematically investigated the employment of both explicit and implicit coordination in a real-life joint activity. Real-life joint activities differ from simple joint actions in that the former are hierarchically structured, composed of events (joint actions) and sub-events (subordinate joint actions) characterized by the achievement of goals and sub-goals (Hard et al., 2006; Hard et al., 2011; Zacks et al., 2001). Investigating the explicit and implicit coordination of joint actions in a real-life joint activity would help decipher how hierarchically related joint actions are coordinated in an efficient way.

To do so, we first need to operationalize the term "coordination". We propose that, to perform a joint action, co-actors must establish a common ground of at least four elements of the joint action: the objects involved (including their identities and properties), the status quo of the task context, the common goal between the co-actors (what changes the joint action would cause to the task context), and division of labor between the co-actors. These elements of the joint actions

in an activity constitute a joint mental model of the activity between the co-actors which is updated whenever necessary as the activity proceeds.

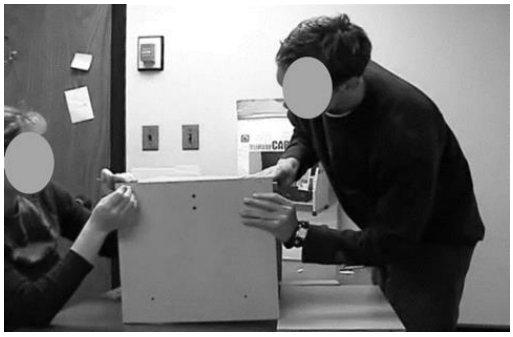

Figure 1: A pair of participants in the middle of the task. Screen shot from the video.

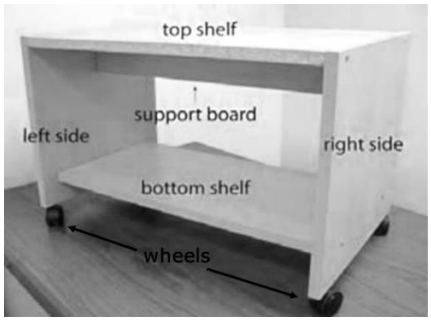

Figure 2: Completed TV cart with its parts labeled, adapted from Heiser et al. (2004).

Following this vein, the present study aimed to investigate how pairs of participants coordinated with each other to assemble a TV cart from its parts (Figure 1). More specifically, we were interested in a) what elements of the joint actions of this activity were coordinated explicitly and implicitly and b) how this happened over time, that is, the temporal pattern of explicit and implicit coordination.

As Figure 2 shows, the TV cart comprised a top shelf, two side boards, one support shelf, one bottom shelf, and four wheels. The participants were given a photo of the completed cart for reference and received no further instruction. Therefore, each pair of participants were free to complete the task in their own idiosyncratic way. The goal of the study was to explore and reveal the common pattern of their coordination.

Towards this end, we coded each pair's utterances and gestures during the task. We also identified the major and subordinate joint actions they performed based on the parts they attached and the behavioral primitives entailed by attaching each part. We then analyzed what elements were explicitly coordinated at what time during the task, and the elements not explicitly coordinated were interpreted as implicitly coordinated. We also analyzed the degree to which the major and subordinate joint actions were coordinated explicitly and implicitly.

Our results suggested a complementary role of explicit and implicit coordination in coordinating a real-life joint activity: Explicit coordination served to identify the affordances of the task context, which allowed co-actors to coordinate the kinematics of their joint actions implicitly. We also argue that the bifurcation of explicit and implicit coordination should be understood on the level of the dyad rather than individuals and that the dyad is the basic unit of joint action.

2.1 Methods

2.1.1 Stimuli

The stimuli were videos previously collected by Heiser and Tversky in 2003 of pairs of unacquainted students as they assembled a TV cart from its parts. The videos were taken by a camera facing the participants' workspace. Figure 1 shows a screen shot from one of the videos featuring a pair of participants.

During the assembly task, each dyad was told to begin with the parts arrayed on the table and use the photo of the completed cart on the package box as a guide; no further instruction was provided. As annotated in Figure 2, the primary parts of the TV cart were: a top shelf, two side boards, one support shelf, one bottom shelf, and four wheels. The boards were joined by screws and pegs, and a screwdriver was available to each dyad.

The study that collected these videos (Heiser et al., 2004) used the assembly task to identify principles of effective instruction design for assembling everyday products but never analyzed the pattern of coordination between the co-actors. The study was approved by the IRB, and the participants were undergraduate students from Stanford University fulfilling a requirement for the introductory psychology class. The current study analyzed 13 out of the 14 videos available. Although all 14 pairs of participants successfully assembled the TV cart, one participant of the excluded pair had an abnormality in his left hand, which rendered the pair's assembling actions systematically different from those of the others. Among the analyzed pairs, six were same-sex dyads (five male and one female), and seven were mixed-sex dyads.

2.1.2 Transcription and Coding

We transcribed the participants' speech, coded its content as well as the gestures and joint actions the participants performed and the ways they coordinated each joint action. For each round of coding, two trained coders coded four out of the 13 videos independently using ELAN (Version 6.0) (2020), and, after attaining a satisfactory inter-rater reliability (kappa value that was above .80 and/or percent agreement that was above .90), one of them continued to code the remaining videos. Disagreement and uncertainty were resolved through discussion. Inter-rater reliability was calculated according to Cohen's (1960) protocol unless otherwise specified. The values of interrater reliability of each round of coding are provided in Appendix C.

2.1.2.1 Speech

2.1.2.1.1 Transcription

Two transcribers transcribed all task-related utterances in each video independently, discussing with each other over uncertain cases. The words and phrases that neither transcriber could identify were marked with placeholders. Sentences with repeated contents were transcribed

and coded as one. In total, we identified 956 single-word/phase sentences and independent and complex clauses, three (0.31%) of which included repeated content, and 30 (3.14%) of which were marked as "untranscribable" as they involved untranscribable parts that rendered the clauses uninterpretable. The untranscribable clauses were excluded from further data analysis.

Single-word/phrase sentences were sentences constituted by a single word or phrase such as "okay", "yes", and "of course". Independent clauses each contained a subject and a predicate and could stand alone as an intelligible sentence, for example, "This (board) goes here." and "We'll put the wheels on last." Complex clauses each contained an independent clause and at least one dependent clause (which could not stand alone), for example, "We can balance this (board) on top after we get these things in." and "It's these things right here, because it's one of the screws." Note that compound clauses, those containing at least two independent sentences joined by conjunctives, were decomposed into independent sentences.

When a clause omitted a grammatical element such as the subject or the verb, we used the verbal and situational context to infer the meaning. For example, the sentence "We can probably slide it in if we…" meant "…if we attach the side board first." We inferred this because the participants went on to attach the side board immediately, which permitted sliding in the aforementioned board. Another example is the sentence "Side board." It could mean "This is a side board." or "That is a side board." depending the speaker's pointing direction. Similarly, based on the speaker's action, "Screw this one." could mean "You screw this one." or "Let's screw this one." These sentences were analyzed as complete sentences. In three cases, the context was occluded so the meaning could not be established—these clauses were excluded from further analysis.

2.1.2.1.2 Coding

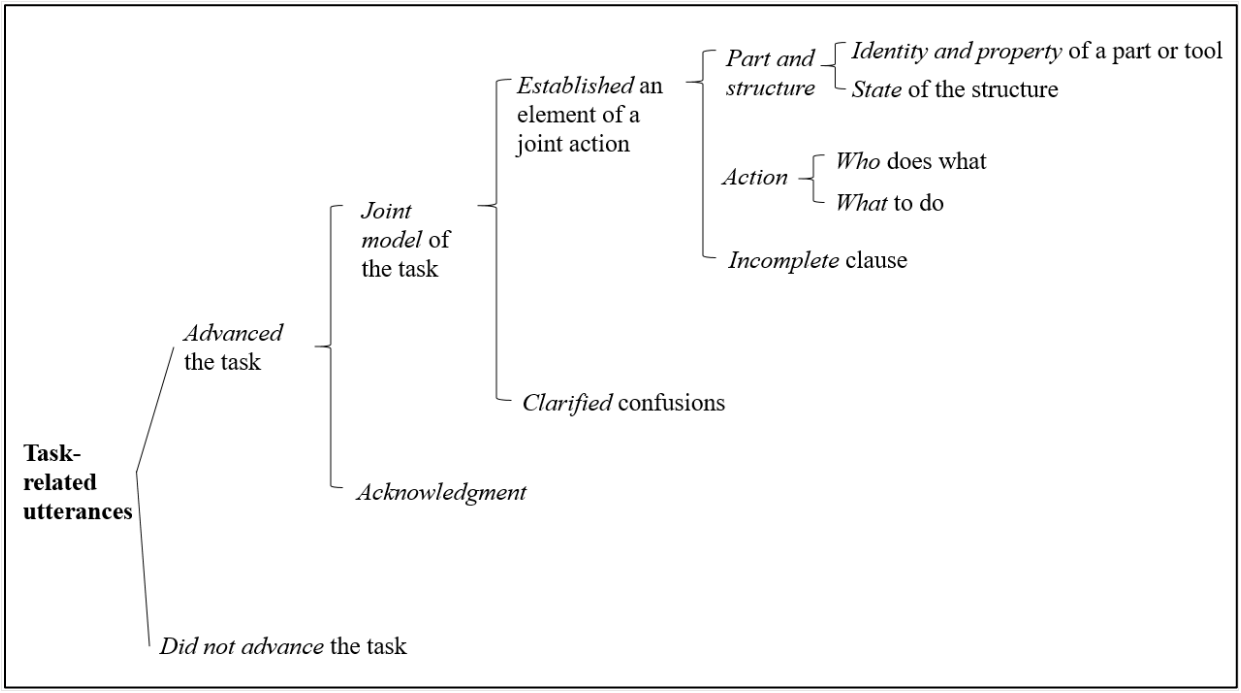

Figure 3: Coding scheme of content of speech. (Italics indicate names of codes.)

Figure 12 shows the hierarchical coding scheme of the utterances by type. We first divided all utterances into those that did and did not advance the participants' understanding of the task (kappa = .711, $p < .001$, percentage agreement = .944). The former included utterances referring to coordination of a joint mental model of the task and acknowledgments (kappa = .837, $p < .001$, percentage agreement = .933). Examples of the latter included utterances irrelevant to the assembling actions (e.g., "Stand up.") and those showing a lack of instantaneous understanding (e.g., "I don't know." and "Let's see.").

We then divided the model-establishing utterances into those that established an element of a joint action (e.g., "This is the bottom." and "You do the screw and I'll try to hold it to get out.") and those that clarified the participants' confusions (e.g., "Is this the top?" and "Why is it [leaning] backwards?") (kappa = .841, $p < .001$, percentage agreement = .940).

The element-establishing utterances were divided into utterances that described the parts and the structure, those that described a specific action, and incomplete clauses (kappa = .821, $p < .001$, percentage agreement = .908). The meanings of the incomplete clauses were completed by their accompanying gestures and/or actions. We separated these utterances out as a unique type of utterances to show even verbal communication might need supplement from other modalities.

We then divided the part-and-structure utterances into those that described the identity and property of a part and those that described the current state of the structure (kappa = .873, $p < .001$, percentage agreement = .953). We also divided the action utterances into those that specified division of labor (i.e., who does what) between the co-actors and those that established the goal (i.e., what to do) of the joint action (kappa = .792, $p < .001$, percentage agreement = .897).

2.1.2.2 Gesture

Table 1: Coding scheme of gesture.

	Code	
Type of Gesture	Deictic Representational Head movement	
Gesture & Speech	Overlapping	Complementary Redundant
	Independent	/

Gesture was defined as an action that was not part of the assembling process except as means of communication (Clark & Krych, 2004). Consistent with McNeill's protocol (1992), the duration of each gesture spanned the preparation, stroke, and retraction phases. As Table 1 shows, we coded the gestures by type and their relationship with the accompanying speech.

Following others, notably Clark and Krych (2004) and McNeill (1992), gestures were coded as *representational*, *deictic*, and *head movement* (kappa = .812, $p <$. 01, percentage agreement = .99). Representational gestures stood for their referents in several possible ways such as pantomiming an action (e.g., lifting and screwing) and referencing a spatial property (e.g., using an upward hand swing to represent the "top" board) (Kita, 2000). Deictic gestures were pointing movements directed at an object or a location, usually performed with a finger, but sometimes by touching, knocking on, or manipulating an object (Clark, 2005). Head movements were head nods or shakes, essentially assent or disagreement, usually serving as feedback.

We then coded whether or not each gesture overlapped speech in content and time (kappa = .876, $p<$.01, percentage agreement = .96). Gestures overlapping speech were either mismatching gestures providing information critical for understanding the speaker's utterances but missing from them (e.g., clarifying "this" and "that" by pointing to the unnamed part or place) or synonymous gestures with the content of speech. These two sub-types of overlapping gestures were further coded as *complementary* and *redundant* respectively (kappa = .852, p < .01, percentage agreement = .88). Gestures that did not overlap speech either occurred without speech or conveyed a meaning irrelevant to the content of speech (e.g., touching one board while talking about another).

2.1.2.3 Joint Action

To assemble the TV cart, the top, support, and bottom shelves needed to be attached to the side boards, and the wheels to the structure. Accordingly, we referred to the actions that attached each pair of parts together as "top-side", "support-side", "bottom-side", and "wheel-structure". Note that "top" is relative to the canonical position of the object with respect to the world, and the "left" and "right" sides are relative to the coder's viewpoint of the object. To

assemble the cart efficiently, the structure-so-far had to be turned and rotated so its actual positions were not the canonical position of the finished object, for example, the "top" board was sometimes on top or bottom, but, to avoid potential confusion, we will refer to the parts with "top", "side", "support", "bottom", and "wheels" as in Figure 2 for the rest of this dissertation.

Table 2: Coding scheme of action.

	Code	
Individual Action	Top-side Support-side Bottom-side Wheel-structure	
Joint Action	Major action	Top-side Support-side Bottom-side Wheel-structure
	Sub-action	Assemble-assemble Assemble-hold Hold-assemble Hold-hold
Direction of Assembly	Doing Undoing	

To code the joint actions, two trained coders first coded which pairs of parts each co-actor attached throughout the task separately (kappa = .919, $p < .01$, percentage agreement = .94). An action began when a participant's hand(s) left the resting position and ended when both hands were withdrawn from the parts or structure. The time periods during which a participant's hands were occluded (by the parts, structure-so-far, or each other) were not considered as part of an action and were excluded from further analysis.

Inter-rater reliability of this round of coding was calculated using Bakeman and Gottman's method (1997): To address the beginning and ending times and type of an action simultaneously, each video was segmented into half-second-long bins. We chose this length to account for the shortest gesture which was slightly over 0.6 second long. The coders coded

which type of action existed in each bin. The beginning and ending times of each action were rounded to match the closer edge of the bin in which they were contained. The bins then formed a list of items whose inter-rater reliability could be calculated as Cohen (1960) prescribed.

Then, joint actions were coded as the time periods during which two co-actors worked on attaching the same parts. Reliability did not apply as joint actions were coded on the basis of the temporal overlap between each two participants' individual actions.

Note that each joint action had its complementary and supplementary parts. *Complementary* referred to when the co-actors' actions were interdependent and served to attain a common goal (e.g., when *A* held two boards in place while *B* screwed them together, Figure 1). *Supplementary* referred to when co-actors' actions were independent, each contributing to the task separately albeit occurring at the same time (e.g., *A* and *B* each attaching a different side board to the bottom shelf). Our investigation focused on the complementary parts of the joint actions as they required more complicated coordination.

We coded the complementary parts of each joint action the same way as the individual actions (kappa = .818, $p < .01$, percentage agreement = .93). Each complementary part of a joint action was referred to as a *major action* for simplicity.

Each major action was decomposed into joint behavioral primitives termed *sub-actions* (Zacks et al., 2001). For example, joining the top board onto a side board included picking up, orienting, holding, inserting (screws), and screwing. Because of their great variety, sub-actions were grouped into two categories, "assemble" and "hold". "Assemble" included any behavioral primitive that was part of putting the parts together (e.g., picking up, orienting, screwing, etc.), and "hold" referred to any motionless action that kept the parts and structure stable in place.

We coded the co-actors' *joint sub-actions* using four codes each corresponding to one pair-wise combination of the sub-actions: "assemble-assemble", "assemble-hold", "hold-assemble", and "hold-hold" (kappa = .853, $p < .01$, percentage agreement = .90); the codes "assemble-hold" and "hold-assemble" reflected the participants' left-right standing positions viewed from an observer's view.

Note that participants made mistakes during the task, in which case they must undo and redo what had been done. Using Bakeman and Gottman's method, we coded the parts of the videos in which participants added more parts to the structure as *doing* actions (regardless of whether those were the correct actions to perform) and the parts that disassembled the structure-so-far as *undoing* actions (kappa = .931, $p < .001$, percentage agreement = .993), and the major actions and sub-actions that fell into the doing and undoing parts of a video were automatically labeled as doing and undoing major and sub-actions.

2.1.2.4 Coordination of Joint Action

To determine how each joint action was coordinated (that is, how their goals and labor division were established), we examined the time between the beginning of the task (because some participants talked about a joint action early on but only executed the action several steps or sub-steps away) and the initiation of each joint action to see if gesture or speech was produced within the interval to coordinate the joint actions. We concluded that, if a joint action proceeded without prior speech or gesture, the action itself was the basis of its coordination.

Joint actions coordinated by speech (and the compound of speech and gesture) were further coded by the nature of speech: *what* for speech that specified what to do next—the common goal—without mentioning labor division, and *who* for speech that specified labor division but not the common goal. Thus, we coded coordination of the major actions and sub-

actions using four codes: “speech-what”, “speech-who”, “gesture”, and “implicitly” (for major actions, kappa = .84, $p < .01$, percentage agreement = .90; for sub-actions, kappa = .97, $p < .01$, percentage agreement = .99).

2.2 Results

Statistical data analysis was conducted using IBM SPSS Statistics for Windows (Version 27.0).

2.2.1 Content of Speech

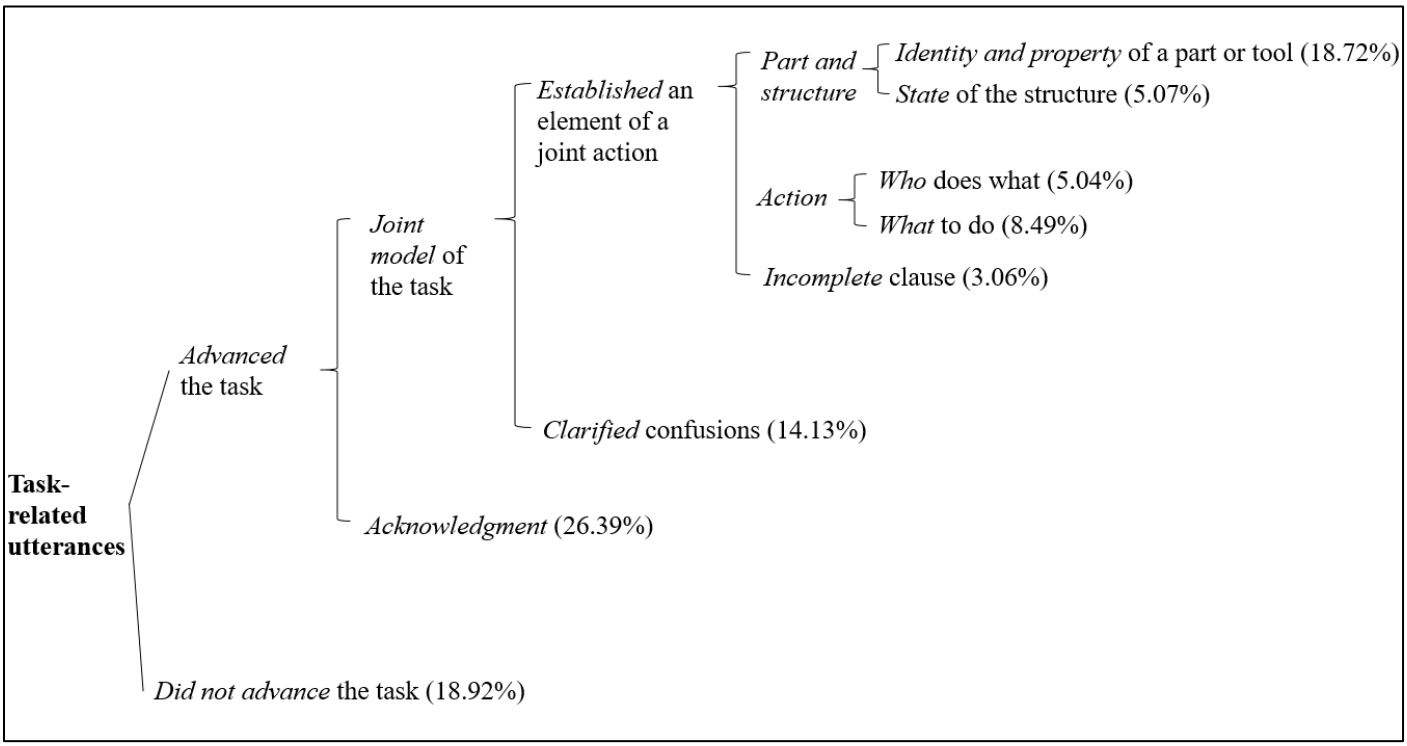

Figure 4: Average rate of each subcategory of utterances.

As in Figure 4, of the total number of task-related utterances averaged over all pairs, 81.08% (*SD* = 8.32%) advanced the participants’ understanding of the task by referring to the specifics of it, and 18.92% (*SD* = 8.32%) did not. The majority (67.45%) of the former (*SD* = 7.33%) served to establish a joint mental model of the task, and 32.55% (*SD* = 7.33%) were acknowledgments. Of the model-establishing utterances, 74.16% (*SD* = 10.01%) specified an

element of a joint action, and 25.84% (*SD* = 10.01%) clarified the co-actors' confusions. Of the element-establishing utterances, 58.83% (*SD* = 10.24%) were about the parts and the structure, 33.37% (*SD* = 11.36%) were about an action, and 7.54% (*SD* = 6.06%) were incomplete sentences whose meanings were supplemented by gestures and/or actions. Of the part-and-structure utterances, 78.48% (*SD* = 12.12%) described the identity or property of a part, and 21.52% (*SD* = 12.12%) described the current state of the structure. Of the action utterances, 62.73% (*SD* = 18.90%) established the goal or sub-goal of the joint action, and 37.27% (SD = 18.90%) specified division of labor between the co-actors.

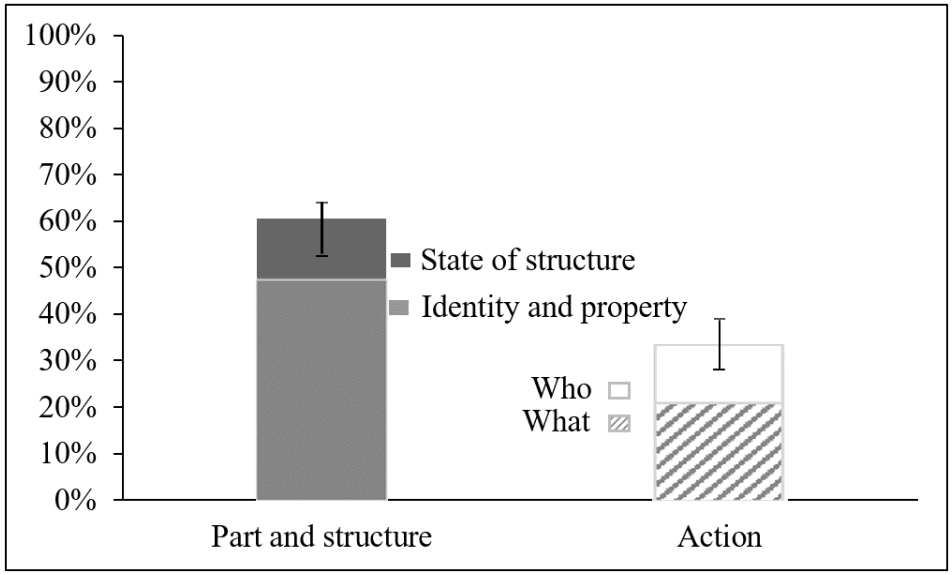

Figure 5: Element-establishing utterances.

As in Figure 5, more of the element-establishing utterances served to describe the parts and structure (i.e., the identity and property of a part or tool and the current state of the structure) than to describe an action (i.e., its goal and labor division) ($t(12) = 4.46$, $p < .001$, $d = .21$).

These results suggested two major roles of speech in coordinating this joint activity: establishing a joint mental model of the affordances of the environmental context and sustaining mutual understanding (through acknowledgment) between the co-actors.

2.2.2 Type of Gesture

We identified 216 gestures from all dyads (M = 16.61, SD = 7.12). The vast majority of the gestures (M = 82.13%, SD = 10.71%) overlapped with speech in content and time. Averaged over all pairs for all gestures, 62.43% (SD = 17.24%) were deictic, 12.87% (SD = 12.36%) were representational, and 17.75% (SD = 14.34%) were head movements.

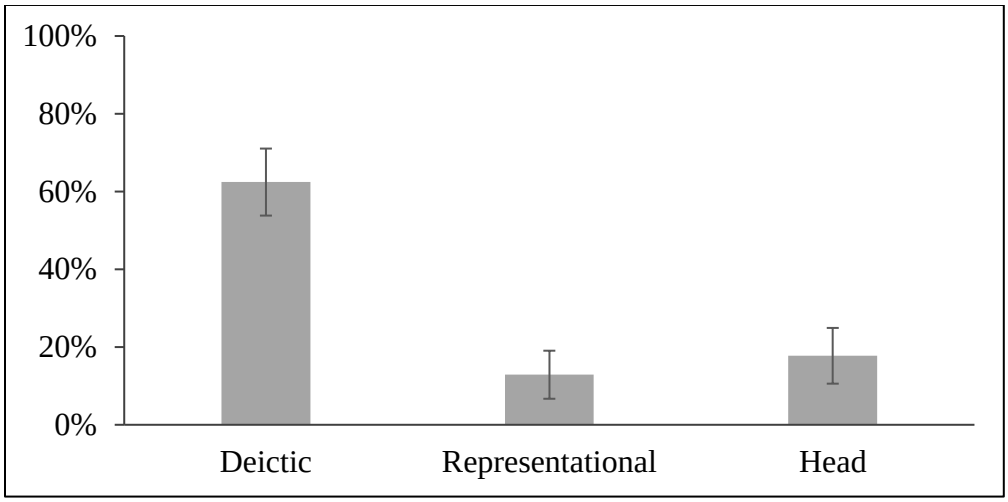

Figure 6: Rate of different gesture types.

As in Figure 6, the rate of gestures was significantly different across gesture type (F(2, 24) = 30.17, p < .001, η^2 = .72). More specifically, deictic gestures occurred more often than both representational gestures (F(1, 12) = 47.07, p < .001, η^2 = .80) and head movements F(1, 12) = 32.06, p < .001, η^2 = .73). These results suggested the primary role of gesture was to guide a co-actor's attention.

2.2.3 Establishment of Joint Mental Model

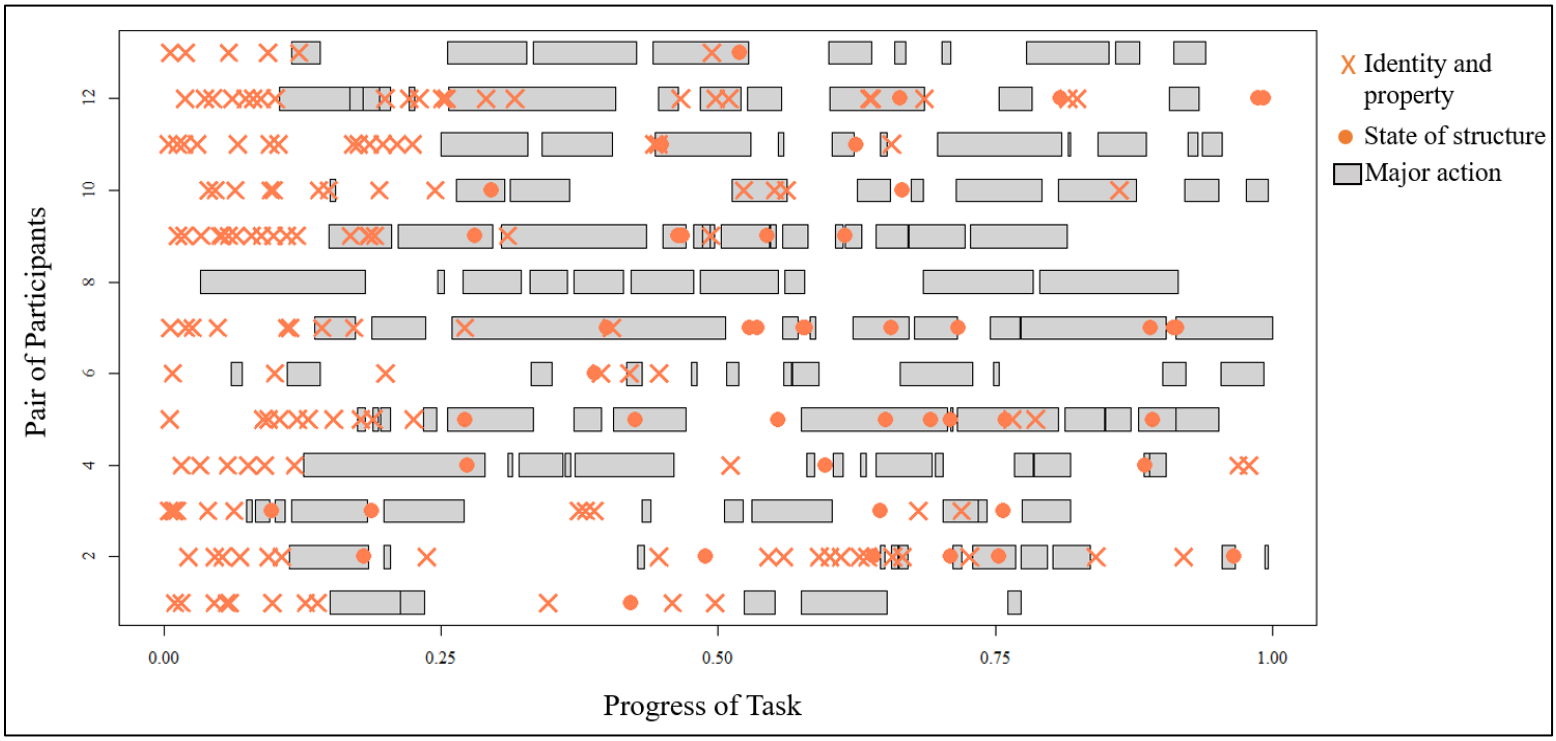

Figure 7: Occurrence of *part-and-structure* utterances and major actions.

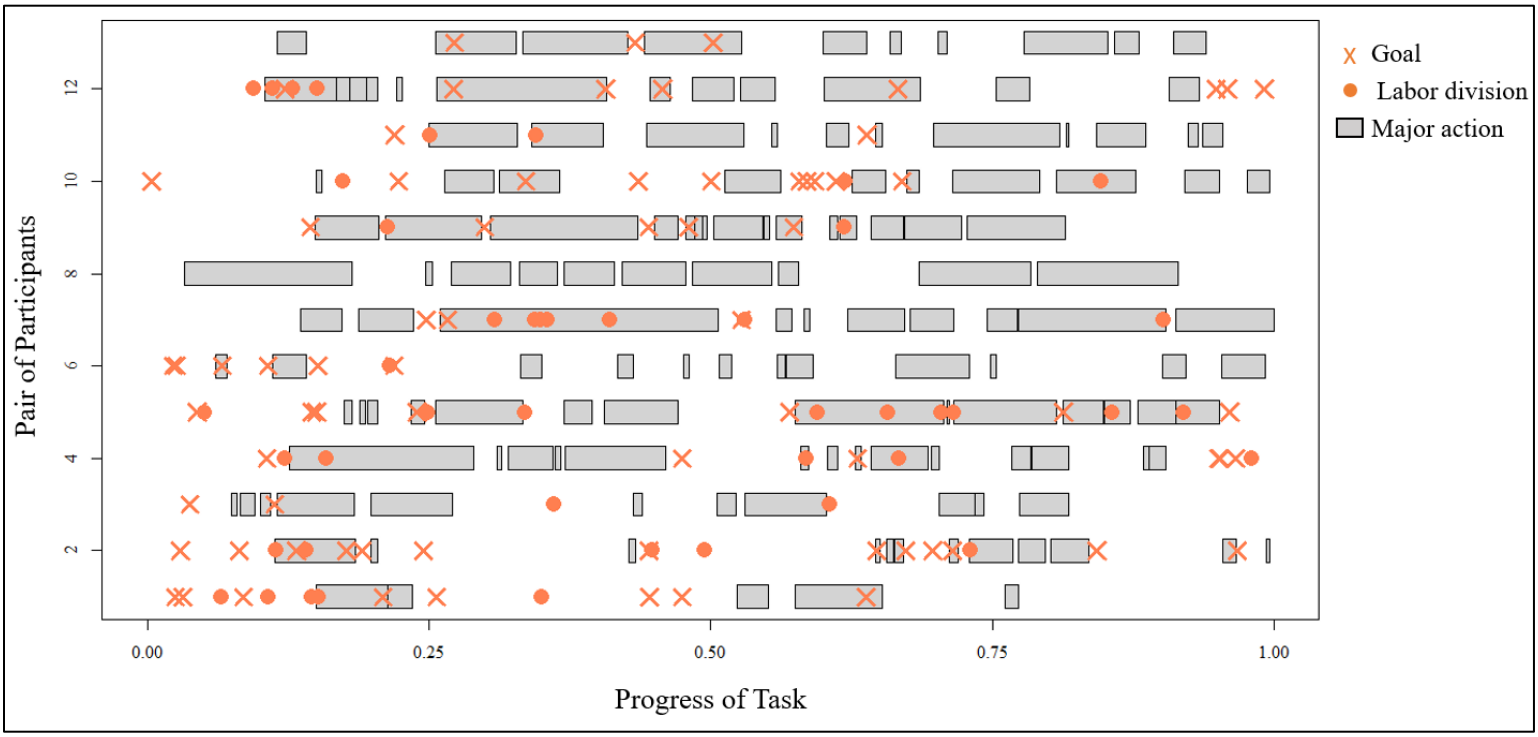

Figure 8: Occurrence of *action* utterances and major actions.

As Figure 7 shows, a significant portion of the part-and-structure utterances occurred at the beginning of the task, and most of them occurred before the participants began their joint assembly. In general, the occurrence of part-and-structure utterances seemed to decline overtime, but the utterances seemed to return in the middle of the task, for example, around half way through the task for Pairs 1, 2, 9, 10, and 12. By contrast, as in Figure 8, action utterances seemed to occur more evenly over time than the part-and-structure utterances. Some action utterances preceded the joint assembly, but this pattern was not as noticeable as in the part-and-structure utterances.

To explain this temporal pattern of the utterances, we defined an attempt at attaching a pair of parts as a major action and an utterance as occurring before or during a major action if the utterance ended between the end times of this major action and the one preceding it. All utterances ending before the end time of the first major action were considered as occurring before or during the first major action.

Then, we compared between the rate of part-and-structure utterances that occurred before or during the participants' initial attempts of attaching a pair of parts and the rate of those that occurred before or during later attempts. We then compared the rate of part-and-structure utterances that occurred before or during later attempts that served as error correction to zero. For the action utterances, we conducted the same comparison between the rates of utterances that occurred before or during the participants' initial and later attempts of attaching a pair of parts. Unlike the part-and-structure utterances, we compared the rate of action utterances that occurred before or during later attempts that served to correct errors to the rate of action utterances that occurred before or during initial attempts that served to correct errors. The results were as follows.

2.2.3.1 Identity, Property, and State

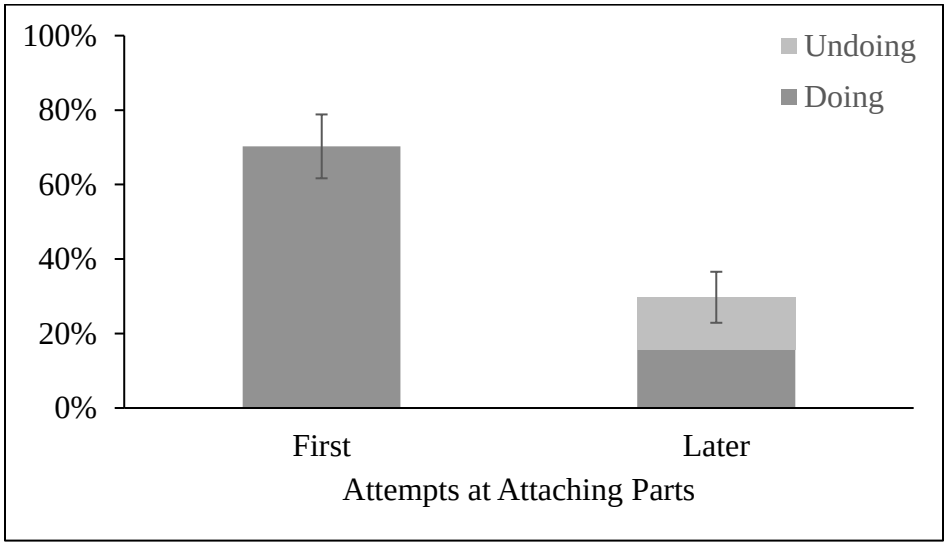

Figure 9: Occurrence of part-and-structure utterances.

As in Figure 9, the majority of the part-and-structure utterances occurred before or during the participants' first joint attempts at attaching a pair of parts, that is, in planning the sequence of major actions prior to or at the beginning of executing it (M = 70.26%, SD = 17.15%), and a few more occurred before or during later attempts (M = 29.74%, SD = 17.15%; $t(12) = 4.26$, $p = .001$, $d = .34$). The proportion of part-and-structure utterances that occurred before or during later attempts that served as error correction (Mdn = 4.35%) was larger than zero ($T = 28.00$, $p = .018$, $r = .47$). These results suggested the part-and-structure utterances established a shared model of the parts early on, and the shared model was sufficient for the entire assembly task unless an error was detected and the shared model must be updated.

2.2.3.2 Common Goal and Division of Labor

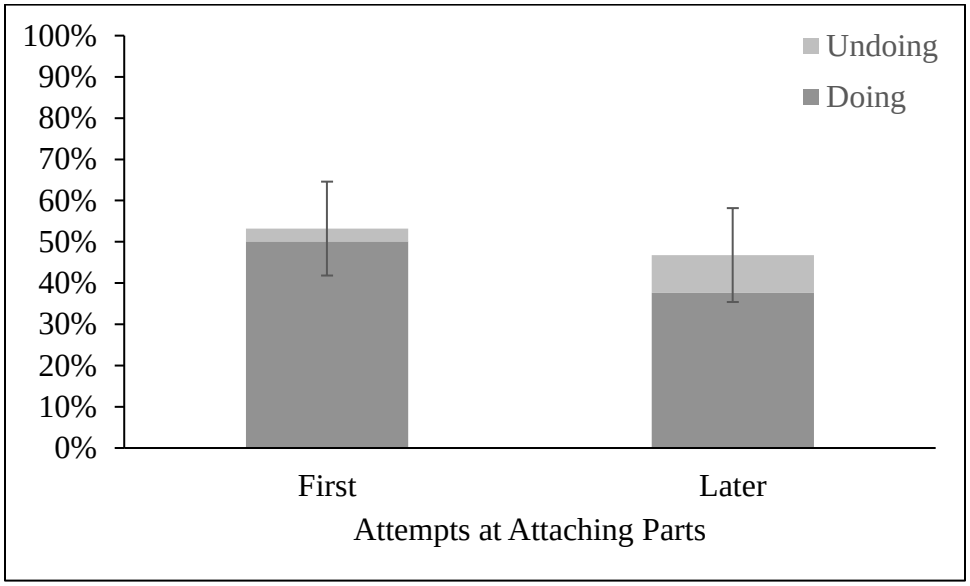

Figure 10: Occurrence of action utterances.

As in Figure 10, the difference between the percent of action utterances occurring before or during the dyads' first (M = 53.21%, SD = 22.77%) and later (M = 46.79%, SD = 22.77%) attempts at attaching a pair of parts was nonsignificant ($t(12) = 0.51$, $p = .62$, $d = .46$). The difference between the percent of action utterances occurring before or during the dyads' first (M = 3.11%, SD = 7.83%) and later (M = 9.17%, SD = 14.81%) error-correcting attempts ($t(12) = -1.18$, $p = .260$, $d = .19$) was also nonsignificant. These results suggested the participants coordinated goals and division of labor based not on progress of the task but on necessity for each joint action.

2.2.4 Coordination of Joint Action

2.2.4.1 Major Action

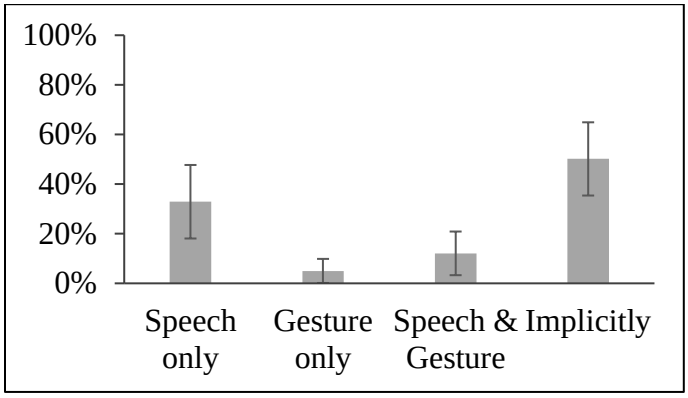

Figure 11: Coordination of major actions.

As in Figure 11, the rate of major actions differed across different means through which they were coordinated ($\chi_F^2(3) = 19.84$, $p < .001$). More specifically, although the difference between the rates of major actions coordinated by speech alone (*Mdn* = 17.67%) and implicitly (*Mdn* = 54.55%) was nonsignificant, the rate of major actions coordinated implicitly was significantly higher than the rates of major actions coordinated by gesture alone (*Mdn* = 0.00%; $p < .001$, $r = .79$) and by both speech and gesture (*Mdn* = 0.00%; $p = .023$, $r = .57$).

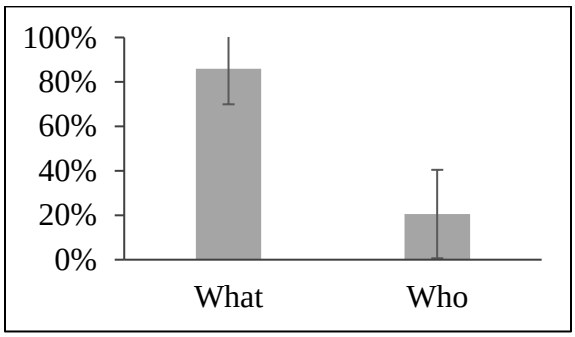

Figure 12: Speech coordinating major actions.

As in Figure 12, of the major actions coordinated by speech and/or gesture, far more referred to a common goal, that is, "what" (*Mdn* = 100.00%), than to division of labor, or "who" (*Mdn* = 0.00%) ($T = 11.00$, $p = .01$, $r = .49$), thus suggesting that more clarification was needed

to specify what needed to be done than who needed to do it. Still, most major actions needed no explicit coordination.

2.2.4.2 Sub-Action

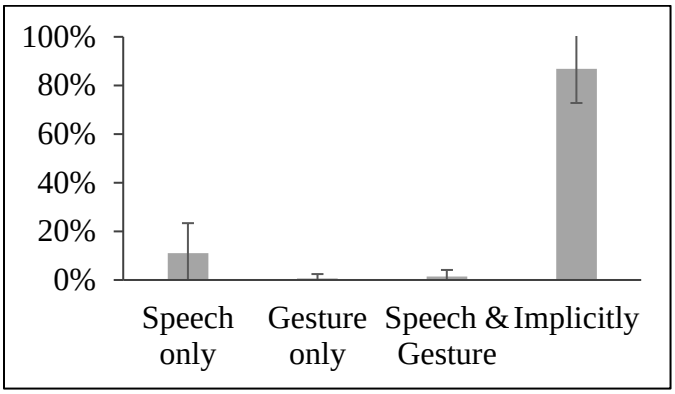

Figure 13: Coordination of sub-actions.

As in Figure 13, the rate of sub-actions differed across different means through which they were coordinated ($\chi_F^2(3) = 33.70$, $p < .001$). More specifically, the rate of sub-actions coordinated implicitly (*Mdn* = 91.89%) was significantly higher than the rates of sub-actions coordinated by gesture alone (*Mdn* = 0.00%; $p < .001$, $r = .92$) and by both speech and gesture (*Mdn* = 0.00%; $p < .001$, $r = .89$). The rate of sub-actions implicitly coordinated was borderline higher than that of sub-actions coordinated by speech alone (*Mdn* = 6.25%; $p = .06$, $r = .51$).

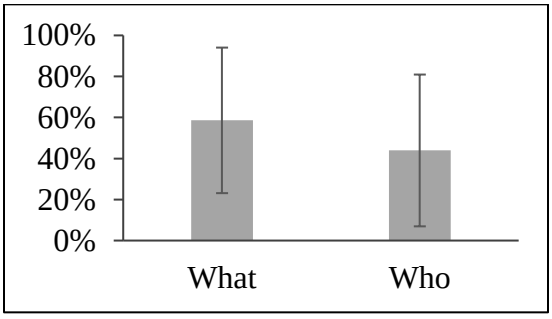

Figure 14: Speech coordinating sub-actions.

As in Figure 14, of the sub-actions coordinated by speech and/or gesture, the difference between common goal (or "what", *M* = 58.58%, *SD* = 35.47%) and division of labor (or "who", *M* = 43.92%, *SD* = 36.99%) was nonsignificant ($t(9) = .64$, $p = .54$, $d = .72$).

3.3 Discussion

In this study, we investigated how pairs of strangers employed explicit and implicit means of coordination when they collaborated to perform a real-life task, assembling a TV cart from its parts. This task usually entails studying and following a set of explicit instructions step-by-step, but participants in our study received no instruction and had to figure the specifics out themselves.

Early on, participants used explicit communicative means, primarily speech but also some gestures to determine how the parts should be configured, that is, which part should go where. These explicit interactions created a shared model of the structure to be completed that was refined and updated as the task progressed. Most dyads continued to specify the properties of the parts as well as the current state of the structure before initiating the first of several major actions needed to attach a pair of major parts. Some major parts (e.g., the side boards) might be attached to several other parts (e.g., the top, support, and bottom boards), and their properties were not repeatedly discussed, nor was the ultimate state of the structure, which resulted in a declining use of speech and gesture over time—except when the partners encountered an error when speech returned to update the local shared model.

The actions needed to create the structure were not so much explicitly specified as the identities, properties, and states of the parts and structure, nor was the order of assembly, even though there were several possible orders. After the shared model of the structure was established, coordination of the actual assembly, that is, the step-by-step actions and the order of actions required to build the object, on both the major and subordinate levels, were improvised as the assembly progressed. Such improvisation was primarily implicit, achieved through observation of the actions themselves and guided by the shared model of the structure and the

current state of assembly. Although the majority of the joint actions were implicitly coordinated, there was a small amount of speech serving as explicit coordination especially for the major actions, the most part of which concerned what needed to be done (i.e., the common goal) rather than who needed to do what (i.e., the division of labor).

Thus, coordination of joint action in a real-life goal-directed activity seems to be the synergy of explicit and implicit coordination with explicit coordination serving primarily to establish and sustain a shared mental model of the environment between the co-actors and implicit coordination serving primarily to specify the specifics (i.e., goal and division of labor) of their actions.

3.3.1 Explicit Coordination: Mental Model of the Environment

3.3.1.1 Prerequisite of Planning

Mattar and Lengyel (2022) defined planning as the selection of an action or a sequence of actions based on their outcomes and the desirability of these outcomes. Planning, therefore, entails estimation as well as evaluation of an action's outcomes. Between estimation and evaluation, one must be able to estimate the outcomes in order to evaluate them, or, in other words, one must establish the association between an action and its outcomes before evaluating the outcomes.

Multiple brain areas have been associated with learning the associations between actions and their outcomes such as the orbitofrontal cortex (OFC), the dorsolateral prefrontal cortex (dlPFC), the ventralmedial prefrontal cortex (vmPFC), and the prelimbic cortex (PL), and the associations seem to be stored in the dorsomedial striatum (DMS) (see Miller & Venditto, 2021, for a review).

However, is it true that *action* is the sole crux of planning, or could we interpret action as the mediator between a stimulus and an outcome or two different states of the same stimulus? A simple example is an action causing changes to a physical environment—in our study, action caused a set of boards to become a structured TV cart. It is therefore possible that, when planning an activity such as assembling a TV cart that involves intense physical interaction with the environment, having a stimulus-outcome association might be the prerequisite for establishing an action-outcome association. In support of this view, stimulus-outcome associations are represented in the brain independently of the action-outcome outcomes: Unlike activity in the dorsal striatum, activity in the ventral striatum correlates not with an action-outcome association but with a stimulus-outcome association (O'Doherty et al., 2004).

The representation of a stimulus-outcome association in a physical environment corresponds to the understanding of how the environment responds to an action (Mattar & Lengyel, 2022). Viewed from a different perspective, when the environment responds to an action, it allows the action to take effect, thus providing an opportunity to act for actors who have the action in their repertoires—what Gibson termed "affordances" of the environment. Therefore, having a mental representation of the environmental affordances allows an actor to make predictions of the outcomes of their actions in the environment, which is the foundation of planning.

In our study, participants actively identified the identities and properties of the parts and the state of the current structure at the beginning of the task and as they began to attach a pair of parts. These identities, properties, and states indicated the affordances of the parts and the structure. For example, the fact that a board was one of the sides of the TV cart indicated it could be attached to three separate boards—the top shelf, the support board, and the bottom shelf. The

two holes on the top of the side board matched the holes on the top board, which indicated the two boards could be joined by screws. Our study did not investigate how the outcome part of the association was activated upon registering the stimuli in the mind, but future research should explore the possibility that representations of stimuli-outcome associations are activated prior to that of their corresponding action-outcome associations. Nevertheless, the established identities, properties, and states of the parts and structure served as elements of a mental model of the task environment for the participants.

3.3.1.2 Planning Together

To jointly plan an activity, co-actors must establish the elements of the mental model on a common ground where they mutually understand the establishment of the elements. Here, we refer to "mutual understanding" as the state of "I know I know, I know you know, and I know you know I know." Such mutual understanding is often achieved through acknowledgment, for example, by providing a confirmatory response. Through acknowledgment, an established element is grounded between the co-actors, and the element is converted from *my* idea into *our* common understanding. In our study, acknowledgment was achieved through speech, gesture, change in action, and silent acceptance. It seemed that, when one of the co-actors proposed to establish an element (i.e., identified an identity, property, or state), the element would be considered as established on the common ground if the other co-actor did not object or raise a question.

With an established common ground, or a shared mental model of the task affordances, co-actors could plan as if they were parts of a single agent. Although participants in our study devoted only a small amount of utterances to planning actions, most of these utterances concerned what needed to be done (by the dyad) rather than who (in the dyad) needed to do it.

That is, they planned the activity based on *our* capacity rather than the capacity of *yours* or *mine*. This result goes in line with Richardson, Marsh, and Baron's study where pairs of participants switched from single to joint lifting of planks of varying sizes based on their combined arm span.

3.3.1.3 Constrained Cognitive Resources

Even when planning jointly, co-actors might make mistakes in the sense that a certain step might not be optimal. In our study, the support board must be attached before the top and bottom shelves to the side boards; otherwise it would be physically impossible to attach the support board. Many pairs of participants did not realize this until they had attached the top or bottom board to the side boards—only then they began to plan for the support board and realized they had made a mistake.

The cognitive resources available for planning seems constrained for normal individuals. Human optimal planning is usually restricted to three to six steps depending on the difficulty of the planning task (for a review, see Mattar & Lengyel, 2022). Snider et al. (2015) conducted a study where a triangular grid of disks of varying sizes scrolled down a touchscreen at a variable speed; the larger the size of the disk, the larger a reward it corresponded to. Participants were tasked with touching one disk in a row at a time, going upwards across the grid, to maximize their accumulated reward. With each disk they touched, they could only continue to touch the two disks adjacent to it in the upper row. The results showed that, for a given task difficulty (i.e., scrolling speed), participants traded off their depth of planning for shorter recalculation periods, thus maintaining their precision of computation.

Such a trade-off between depth of planning and recalculation periods could be a strategy defending against the risk of executing the plan as the world changes. Extrinsic uncertainty during planning could exist both independently of an agent's actions and because the agent is

unsure how the world might respond to their actions. In both cases, the agent lacks the knowledge to plan far ahead, and, instead of pre-planning a sequence of actions and adhering to it dogmatically, it might be more helpful to compute the best actions within a certain limit and preserve the flexibility to choose future actions based on future "current" states.

3.3.2 Implicit Coordination: Specifics of an Action

As mentioned above, in our study, a mental model of the environmental affordances facilitated the co-actors' conceptualization of ways to complete the task and helped narrow down the possible sequences to attach the parts, but the exact sequences of the major actions and sub-actions seemed determined from the bottom up through division of labor of the sub-actions.

Implicit division of labor of a sub-action seemed achieved through rapid intention detection. The human eye is sensitive to intentions embedded in action. Partners code each other's actions in terms of both what they are (e.g., a grasp) and why they are performed (e.g., grasping to remove) (Uithol et al., 2011). Humans could distinguish between different intentions of the same action (e.g., grasping to drink vs. grasping to pour from a water bottle) based solely on kinematic information such as wrist height and movement of the dorsum plane (Cavallo et al., 2016). Such subtle differences in kinematic and intentional information have been found to elicit different patterns of brain activity during viewing (Koul et al., 2018).

Within the context of TV-cart assembling, co-actors should be able to recognize their partners' sub-actions (e.g., picking up the top board) and detect their subsequent intentions (e.g., picking up the top board to orient it to a side board) based on each other's early-stage body movements, which would allow them to adopt each other's intentions as their own (thus establishing a common sub-goal) and flexibly complement each other's actions accordingly (thus establishing division of labor). In such cases, the co-actor who initiated the sub-action also

preemptively adopted a role, which would further restrict the other's actions. For example, when attaching the top board to the side board, if one co-actor opted to *hold* the boards, the only option left for the other would be to *insert the screws*. Furthermore, the intentions of the initial sub-actions (e.g., picking up and orienting the top board and a side board) would determine the goal of the corresponding major action (e.g., attaching the top board to the side board). Thus, intention detection seemed to enable the establishment of division of labor and goal of the sub-actions, which, in turn, would pinpoint the sequence and goal of the major actions among all possibilities.

In this sense, if we were to ultimately consider *collaboration* as achieving a common goal through appropriate coordination of labor division between co-actors, action itself should be considered as an efficient means of coordination and part of the "conversation" between co-actors. In support of this view, research has shown that co-actors could achieve efficient division of labor silently over rounds of iterated joint processes (e.g., uncovering tiles to determine the existence of a potentially hidden object, Andrade-Lotero & Goldstone, 2021; joint shepherding, Nalepka et al., 2017) as they observe and learn about each other's strategies and behavioral biases. Some of these behavioral biases become stabilized over time and are admitted as norms or conventions of collaboration (Ho et al., 2016) which each co-actor follow and expect the other to follow.

3.3.3 Sporadic Explicit Coordination of Action Specifics

Although the cases were rare, we must note that some utterances in the assembling task did coordinate goals and division of labor of the actions. Based on our inspection, oral coordination of goals and division of labor occurred frequently when one co-actor was visually inaccessible to the other (e.g., when one of the co-actors was occluded by the structure). Visual

accessibility has been found to affect co-actors' quality and mode of coordination (e.g., Clark & Krych , 2004; Richardson, Marsh, Isenhower, et al., 2007). We therefore suggest that future studies should investigate coordination of joint action in non-face-to-face activities such as remote surgery and team sports in which joint attention is hard to achieve.

One point worth noting is that the utterances serving to coordinate goals were primarily concerned with the goal of a major action as if the sub-goals of the sub-actions would automatically follow once the major goal was established. Dezfouli and Balleine (2013) proposed the idea of *chunking* in planning and habit formation; actions that consistently follow from one another are "chunked" and evaluated as a unit during planning, which might be one of the mechanisms through which habits are formed. Viewed from the opposite direction, when co-actors share a habit of action, or when the objects to be acted upon suggest a habit of action, co-actors might as well plan their joint actions based on chunks of habitual behaviors. Then, might there be a basic unit of planning in planning real-life events? Might planning real-life events correspond to the hierarchical structure of perception and memory of events as punctuated by goals and sub-goals? These questions are also worth exploring in future research.

3.3.4 Application in Human-Computer Interaction

The idea of a bifurcation of explicit and implicit communication in collaboration has also been explored in human-computer interaction.

For example, Breazeal et al. (2005) studied the impact of explicit and implicit non-verbal social cues on the performance of human-robot collaboration. Similar to our definition, the authors defined explicit communication as deliberately sending messages to share specific information and implicit communication as conveying information inherent in behavior but not deliberately communicated. In the study, a human participant teamed up with a humanoid robot

to perform four interactive tasks: teaching the robot the name and locations of the three buttons in its front; checking if the robot has acquired such knowledge; having the robot turn on all buttons; and telling the robot that the "all-the-buttons-on task" is done.

Compared to when the robot could only communicate explicitly, that is, responding to the human's direct questions with head nods and shakes, the possibility to perform implicit communication, that is, to communicate its attentional states through changes in its gaze direction and shrugging and making questioning facial expressions to suggest confusion, significantly increased their efficiency of collaboration with the human. The human partners took shorter lengths of time to complete the task and were better able to detect and mitigate robot errors. The authors argued that the human participants built a mental model of the attentional and mental states of the robot during collaboration the way they would interacting with a human collaborator, and the implicit cues facilitated the establishment of this mental model.

One point worth noting in Breazeal et al.'s study is the human dominated the human-robot collaboration and that the collaborative task was pre-planned. In recent years, more studies have been conducted exploring mutual online adjustment between a human and a robot. For example, Che et al. (2020) investigated whether the robot's proactive communication of its intention could facilitate collision avoidance and efficient trajectory planning when a robot and a human crossed paths during navigation. The authors introduced three conditions of human-robot communication in this study. In the *explicit and implicit* condition, the robot explicitly communicated its intent to the human (i.e., to let the human go first or to take priority) through a wearable vibration motor. It also sent implicit information through changes in its moving speed and direction based on its prediction of the human's movement. In the *implicit only* condition, the robot only changed its speed and direction and did not send any haptic signal. In the *baseline*

condition, the robot did not predict the human's movements and simply performed collision avoidance.

The results showed explicit haptic communication clarified the robot's intentions for the human during navigation, which also led to shorter path lengths for the human and higher levels of trust in the robot. Implicit communication had a similar yet milder effect compared to when no prediction or communication was performed. Thus, it seemed the robot's proactive communication with the human, be it explicit or implicit, facilitated the efficiency of human-robot collaboration. This could be because such communication provided the human a way of interaction with the robot that was akin to human-human interaction—the competence of which human-robot collaboration aspires to achieve.

In terms of efficient means of human-robot interaction, Mahadevan et al. (2021) compared between multiple explicit and implicit means of human-robot labor division in a VR environment. In this task, the human participant and a robotic arm paired up to complete a block stacking task where the human was responsible for the yellow and the robot was responsible for the black blocks. The robot was not informed in advance that it could only move and manipulate the black blocks and must learn this rule from its interaction with the human. The authors compared four explicit means through which the human allocated different blocks to themselves and the robot: *voice*, verbally ordering the robot to maneuver a certain block; *menu*, selecting allocation of a block through a pop-up menu; *subtle relocation*, allocating a block through pushing (to the robot's side) and pulling (to their own side); and *fixed territories*, moving a block to pre-determined territories that belonged to the human, the robot, and the group. The four implicit means required the robot to infer about block allocation through implicit cues such as a block's distance to its last action, its base, and the human's gazing direction. The results showed

the implicit means resulted in comparable performance of human-robot collaboration to that of *voice*, and they outperformed the other explicit means. This suggested implicit human-robot collaboration gave the robot the capacity to infer the human's intentions as well as rules of the collaboration, thus freeing the human from explicit division of labor and promoting efficiency of collaboration.

As delightful as it is to see evidence of explicit and implicit communication in facilitating human-robot collaboration, it must be noted that human-robot collaboration is far from reaching the level of human-human collaboration.

First and foremost, most studies of human-robot interaction focus on simple collaborative tasks featuring one robot and one human collaborating on a one-stop action (i.e., not a hierarchical event where actions are inherently interconnected and serve a hierarchy of goals). In real life, collaborative tasks could take multiple agents coordinating with one another at the same time (e.g., navigating streets packed with pedestrians and vehicles), and they could be so complicated that they must be achieved step-by-step through the achievement of goals and sub-goals (e.g., cooking, assembling, paying for groceries, moving internationally). Future research should investigate the pattern of communication during multi-human interaction with single or multiple robots in more complicated tasks.

Second, although current models of human behaviors (e.g., Breazeal et al., 2005; Che et al., 2020) do take environmental factors into consideration, they might suffer from oversimplifying the role of the environment in the model. For example, Breazeal et al., (2005) equated understanding of the environment to connecting an object with its name. This could be problematic as any planning in the physical environment would require knowledge of the connection between the objects involved, their properties, and their affordances. Future research

should consider integrating such association chains into their predictive models of human behaviors.

Third, humans tend to switch flexibly among different explicit and implicit means of communication during collaboration. Research on training robots to understand the human's intentionality from a mixture of means of communication seems to be lacking. Mahadevan et al. (2021) compared between the effectiveness of different means of communication, but participants in their study could only apply one communicative means each time they interacted with the robot, which could lead to the robot's misinterpretation of the human's intention and rule of the collaboration. For example, the robot could consider the human as circumscribing a territory (i.e., allocating every block within the territory to themselves) with their eye gaze while the human might actually be considering possible movements of a block. It is important to train collaborative robots in more naturalistic collaborative situations with more naturalistic stimuli from the human.

In general, there seems to be a spectrum of the degree of human-vs.-robot dominance in human-robot collaboration. When a human or a robot dominates the collaboration, they are often responsible for monitoring and adjusting to the other, and the human and the robot could mutually adjust to each other in the middle of the spectrum. Different collaborative tasks might require different degrees of human/robot dominance. For example, when robot recognition of human gestures and language is imperfect, humans might serve best to detect robot mistakes and minimize collaborative errors. In cases where humans suffer from limited cognitive resources such as multi-step planning and planning during emergency, it might be best for AI systems to process the information, make the plan, and instruct humans' actions. As Mahadevan et al. (2021) mentioned, human collaborators prefer to work with proactive robots but also sometimes

wish to exert more careful control over the robot. Then the question is, how to enable robots to fluidly switch between different modes of collaboration with humans? This is another question awaiting answers from further research.

Conclusion

Through coding and qualitative as well as quantitative analysis of the use of speech, gesture, and action in coordinating major and subordinate joint actions in a real-life activity, we found explicit means of coordination such as speech and gesture served primarily to establish and sustain a shared mental model of the environmental affordances between the co-actors. The goals and division of labor of the actions were primarily coordinated implicitly using action itself, which was scaffolded by the shared model of the environment. Future directions of the role of the mental model in planning as well as applications of human-human coordination in human-computer coordination was suggested.

References

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nature Reviews Neuroscience*, *7*(4), 268-277. https://doi.org/10.1038/nrn1884

Andrade-Lotero, E., & Goldstone, R. L. (2021). Self-organized division of cognitive labor. *PloS one*, *16*(7), e0254532. https://doi.org/10.1371/journal.pone.0254532

Bakeman, R., & Adamson, L. B. (1984). Coordinating attention to people and objects in mother-infant and peer-infant interaction. *Child Development*, *55*(4), 1278-1289.

Bakeman, R., & Gottman, J. M. (1997). *Observing interaction: An introduction to sequential analysis.* Cambridge University Press. https://doi.org/10.1017/CBO9780511527685

Becchio, C., Sartori, L., Bulgheroni, M., & Castiello, U. (2008). Both your intention and mine are reflected in the kinematics of my reach-to-grasp movement. *Cognition*, *106*(2), 894-912. https://doi.org/10.1016/j.cognition.2007.05.004

Becchio, C., Sartori, L., & Castiello, U. (2010). Toward you: The social side of actions. *Current Directions in Psychological Science*, *19*(3), 183-188. https://doi.org/10.1177/0963721410370131

Bekkering, H., De Bruijn, E. R., Cuijpers, R. H., Newman-Norlund, R., Van Schie, H. T., & Meulenbroek, R. (2009). Joint action: Neurocognitive mechanisms supporting human interaction. *Topics in Cognitive Science*, *1*(2), 340-352. https://doi.org/10.1111/j.1756-8765.2009.01023.x

Brass, M., Bekkering, H., & Prinz, W. (2001). Movement observation affects movement execution in a simple response task. *Acta Psychologica*, *106*(1-2), 3-22. https://doi.org/10.1016/S0001-6918(00)00024-X

Breazeal, C., Kidd, C. D., Thomaz, A. L., Hoffman, G., & Berlin, M. (2005). Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *2005 IEEE/RSJ international conference on intelligent robots and systems* (pp. 708-713). IEEE.

Buccino, G., Binkofski, F., Fink, G. R., Fadiga, L., Fogassi, L., Gallese, V., ... & Freund, H. J. (2001). Action observation activates premotor and parietal areas in a somatotopic manner: An fMRI study. *European Journal of Neuroscience*, *13*(2), 400-404. https://doi.org/10.1111/j.1460-9568.2001.01385.x

Call, J., & Tomasello, M. (2007). Comparing the gestures of apes and monkeys. In J. Call & M. Tomasello (Eds.), *The Gestural Communication of Apes and Monkeys* (pp. 197–220). Taylor & Francis Group/Lawrence Erlbaum Associates.

Calvo-Merino, B., Glaser, D. E., Grèzes, J., Passingham, R. E., & Haggard, P. (2005). Action observation and acquired motor skills: An fMRI study with expert dancers. *Cerebral Cortex, 15*(8), 1243-1249. https://doi.org/10.1093/cercor/bhi007

Candidi, M., Curioni, A., Donnarumma, F., Sacheli, L. M., & Pezzulo, G. (2015). Interactional leader–follower sensorimotor communication strategies during repetitive joint actions. *Journal of the Royal Society Interface*, *12*(110), 20150644. https://doi.org/10.1098/rsif.2015.0644

Castiello, U., Lusher, D., Mari, M., Edwards, M. G., & Humphreys, G. W. (2002). Observing a human or a robotic hand grasping an object: Differential motor priming effects. In W. Prinz & B. Hommel (Eds.), *Attention and Performance XIX* (pp. 314–334). MIT Press.

Cavallo, A., Koul, A., Ansuini, C., Capozzi, F., & Becchio, C. (2016). Decoding intentions from movement kinematics. *Scientific Reports, 6*(1), 37036. https://doi.org/10.1038/srep37036

Chaminade, T., Marchant, J. L., Kilner, J., & Frith, C. D. (2012). An fMRI study of joint action–varying levels of cooperation correlates with activity in control networks. *Frontiers in Human Neuroscience, 6*, 179. https://doi.org/10.3389/fnhum.2012.00179

Che, Y., Okamura, A. M., & Sadigh, D. (2020). Efficient and trustworthy social navigation via explicit and implicit robot–human communication. *IEEE Transactions on Robotics*, *36*(3), 692-707. https://doi.org/10.1109/tro.2020.2964824

Clark, H. H. (1996). *Using Language.* Cambridge University Press.

Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language, 50*(1), 62-81. https://doi.org/10.1016/j.jml.2003.08.004

Clark, H. H. (2005). Coordinating with each other in a material world. *Discourse Studies*, *7*(4–5), 507-525. https://doi.org/10.1177/1461445605054404

Clark, H. H., & Schaefer, E. F. (1989). Contributing to discourse. *Cognitive Science, 13*(2), 259-294. https://doi.org/10.1016/0364-0213(89)90008-6

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition, 22*(1), 1-39. https://doi.org/10.1016/0010-0277(86)90010-7

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37-46. https://doi.org/10.1177/001316446002000104

Decety, J., Grezes, J., Costes, N., Perani, D., Jeannerod, M., Procyk, E., ... & Fazio, F. (1997). Brain activity during observation of actions. Influence of action content and subject's strategy. *Brain: A Journal of Neurology*, *120*(10), 1763-1777. https://doi.org/10.1093/brain/120.10.1763

Dezfouli, A., & Balleine, B. W. (2013). Actions, action sequences and habits: evidence that goal-directed and habitual action control are hierarchically organized. *PLoS Computational Biology*, *9*(12), e1003364. https://doi.org/10.1371/journal.pcbi.1003364

di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V. & Rizzolatti, G. (1992). Understanding motor events: a neurophysiological study. *Experimental Brain Research, 91*(1), 176–180. https://doi.org/10.1007/BF00230027

Driskell, J. E., & Radtke, P. H. (2003). The effect of gesture on speech production and comprehension. *Human Factors*, *45*(3), 445-454. https://doi.org/10.1518/hfes.45.3.445.27258

Edwards, M. G., Humphreys, G. W., & Castiello, U. (2003). Motor facilitation following action observation: A behavioural study in prehensile action. *Brain and Cognition, 53*(3), 495-502. https://doi.org/10.1016/S0278-2626(03)00210-0

ELAN (Version 6.0) [Computer software]. (2020). Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Retrieved from https://archive.mpi.nl/tla/elan

Ekman, P. and Friesen,W.V. (1972). Hand movements. *Journal of Communication*, *22*(4), 353–374. https://doi.org/10.1111/j.1460-2466.1972.tb00163.x

Flom, R., Lee, K. & Muir, D. (Eds.). (2006). *Gaze-following: Its development and significance.* Lawrence Erlbaum Associates.

Fogassi, L., Ferrari, P. F., Gesierich, B., Rozzi, S., Chersi, F., & Rizzolatti, G. (2005). Parietal lobe: From action organization to intention understanding. *Science, 308*(5722), 662–666. https://doi.org/10.1126/science.1106138

Frick-Horbury, D., & Guttentag, R. E. (1998). The effects of restricting hand gesture production on lexical retrieval and free recall. *The American Journal of Psychology, 111*(1), 43-62. https://doi.org/10.2307/1423536

Frith, C. D., Frith, U. (2008). Implicit and explicit processes in social cognition. *Neuron, 60*(3), 503–510. https://doi.org/10.1016/j.neuron.2008.10.032

Gallese, V., Fadiga, L., Fogassi, L. & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain, 119*(2), 593–609. https://doi.org/10.1093/brain/119.2.593

Gangopadhyay, N., & Schilbach, L. (2012). Seeing minds: a neurophilosophical investigation of the role of perception-action coupling in social perception. *Social Neuroscience, 7*(4), 410-423. https://doi.org/10.1080/17470919.2011.633754

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy?. *Trends in Cognitive Sciences*, 8(1), 8-11. https://doi.org/10.1016/j.tics.2003.10.016

Garrod S., & Pickering, M. J. (2009). Joint action, interactive alignment, and dialog. *Topics in Cognitive Science, 1*(2), 292–304. https://doi.org/10.1111/j.1756-8765.2009.01020.x

Georgiou, I., Becchio, C., Glover, S., & Castiello, U. (2007). Different action patterns for cooperative and competitive behaviour. *Cognition, 102*(3), 415-433. https://doi.org/10.1016/j.cognition.2006.01.008

Gibson, J. J. (1979/2014). *The ecological approach to visual perception: Classic edition.* Psychology Press. https://doi.org/10.4324/9781315740218

Goebl, W., & Palmer, C. (2009). Synchronization of timing and motion among performing musicians. *Music Perception, 26*(5), 427-438. https://doi.org/10.1525/mp.2009.26.5.427

Goldin-Meadow, S. (2003). *Hearing gesture: How our hands help us think.* Belknap Press of Harvard University Press.

Goldin-Meadow, S., Nusbaum, H., Kelly, S. D., & Wagner, S. (2001). Explaining math: Gesturing lightens the load. *Psychological Science, 12*(6), 516-522. https://doi.org/10.1111/1467-9280.00395

Goldin-Meadow, S., & Wagner, S. M. (2005). How our hands help us learn. *Trends in Cognitive Sciences, 9*(5), 234-241. https://doi.org/10.1016/j.tics.2005.03.006

Goupil, L., Wolf, T., Saint-Germier, P., Aucouturier, J. J., & Canonne, C. (2021). Emergent shared intentions support coordination during collective musical improvisations. *Cognitive Science, 45*(1), e12932. https://doi.org/10.1111/cogs.12932

Grèzes, J., Tucker, M., Armony, J., Ellis, R., & Passingham, R. E. (2003). Objects automatically potentiate action: An fMRI study of implicit processing. *European Journal of Neuroscience, 17*(12), 2735-2740. https://doi.org/10.1046/j.1460-9568.2003.02695.x

Hard, B. M., Recchia, G., & Tversky, B. (2011). The shape of action. *Journal of Experimental Psychology: General, 140*(4), 586. https://doi.org/10.1037/a0024310

Hard, B. M., Tversky, B., & Lang, D. S. (2006). Making sense of abstract events: Building event schemas. *Memory & Cognition, 34*(6), 1221-1235. https://doi.org/10.3758/BF03193267

Hay, D. F. (1979). Cooperative interactions and sharing between very young children and their parents. *Developmental Psychology, 15*(6), 647. https://doi.org/10.1037/0012-1649.15.6.647

Hay, D. F., & Murray, P. (1982). Giving and requesting: Social facilitation of infants' offers to adults. *Infant Behavior and Development, 5*(2-4), 301-310. https://doi.org/10.1016/S0163-6383(82)80039-8

Heiser, J., Phan, D., Agrawala, M., Tversky, B., & Hanrahan, P. (2004). Identification and validation of cognitive design principles for automated generation of assembly instructions. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, 311-319. https://doi.org/10.1145/989863.989917

Heylighen, F. (2016). Stigmergy as a universal coordination mechanism I: Definition and components. *Cognitive Systems Research, 38*, 4-13. https://doi.org/10.1016/j.cogsys.2015.12.002

Ho, M. K., MacGlashan, J., Greenwald, A., Littman, M. L., Hilliard, E., Trimbach, C., ... & Austerweil, J. L. (2016). Feature-based joint planning and norm learning in collaborative games. *Proceedings of the 38th Annual Conference of the Cognitive Science Society,* 1158–1163.

Horschler, D. J., MacLean, E. L., & Santos, L. R. (2020). Do non-human primates really represent others' beliefs?. *Trends in Cognitive Sciences, 24*(8), 594-605. https://doi.org/10.1016/j.tics.2020.05.009

Hostetter, A. B. (2011). When do gestures communicate? A meta-analysis. *Psychological Bulletin, 137*(2), 297. https://doi.org/10.1037/a0022128

Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J. C., & Rizzolatti, G. (2005). Grasping the intentions of others with one's own mirror neuron system. *PLoS Biology, 3*(3), e79. https://doi.org/10.1371/journal.pbio.0030079

Jacob, P. (2008). What do mirror neurons contribute to human social cognition?. *Mind & Language, 23*(2), 190–223. https://doi.org/10.1111/j.1468-0017.2007.00337.x

Jacobs, N., & Garnham, A. (2007). The role of conversational hand gestures in a narrative task. *Journal of Memory and Language, 56*(2), 291-303. https://doi.org/10.1016/j.jml.2006.07.011

Jellema, T., Baker, C. I., Wicker, B., & Perrett, D. I. (2000). Neural representation for the perception of the intentionality of actions. *Brain and Cognition, 44*(2), 280-302. https://doi.org/10.1006/brcg.2000.1231

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics, 14*(2), 201-211. https://doi.org/10.3758/BF03212378

Kang, S. and Tversky, B. (2016). From hands to minds: Gestures promote understanding. *Cognitive Research: Principles and Implications*, *1*(1), 4. https://doi.org/10.1186/s41235-016-0004-9

Kelly, S. D., Kravitz, C., & Hopkins, M. (2004). Neural correlates of bimodal speech and gesture comprehension. *Brain and Language, 89*(1), 253-260. https://doi.org/10.1016/S0093-934X(03)00335-3

Kita, S. (2000). How representational gestures help speaking. In D. McNeill (Ed.), *Language and gesture* (pp. 162-185). Cambridge University Press.

Knoblich, G., & Jordan, J. S. (2003). Action coordination in groups and individuals: Learning anticipatory control. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(5), 1006. https://doi.org/10.1037/0278-7393.29.5.1006

Knoblich, G., & Sebanz, N. (2008). Evolving intentions for social interaction: from entrainment to joint action. *Philosophical Transactions of the Royal Society B: Biological Sciences, 363*(1499), 2021-2031. https://doi.org/10.1098/rstb.2008.0006

Kobayashi, H., Yasuda, T., Igarashi, H., & Suzuki, S. (2018). Language use in joint action: The means of referring expressions. *International Journal of Social Robotics*, *12*, 1021–1029. https://doi.org/10.1007/s12369-017-0462-3

Koul, A., Cavallo, A., Cauda, F., Costa, T., Diano, M., Pontil, M., & Becchio, C. (2018). Action observation areas represent intentions from subtle kinematic features. *Cerebral Cortex, 28*(7), 2647-2654. https://doi.org/10.1093/cercor/bhy098

Kourtis, D., Sebanz, N., & Knoblich, G. (2013). Predictive representation of other people's actions in joint action planning: An EEG study. *Social Neuroscience, 8*, 31-42. https://doi.org/10.1080/17470919.2012.694823

Krauss, R. M. (1998). Why do we gesture when we speak?. *Current Directions in Psychological Science, 7*(2), 54-54. https://doi.org/10.1111/1467-8721.ep13175642

Krauss, R. M., Dushay, R. A., Chen, Y., & Rauscher, F. (1995). The communicative value of conversational hand gestures. *Journal of Experimental Social Psychology*, *31*, 533-552. https://doi.org/10.1006/jesp.1995.1024

Lakin, J. L., & Chartrand, T. L. (2003). Using nonconscious behavioral mimicry to create affiliation and rapport. *Psychological Science, 14*(4), 334-339. https://doi.org/10.1111/1467-9280.14481

Levelt, W. J. M. (1989). *Speaking: From intention to articulation.* MIT Press. https://doi.org/10.2307/1423219

Mahadevan, K., Sousa, M., Tang, A., & Grossman, T. (2021). "Grip-that-there": An investigation of explicit and implicit task allocation techniques for human-robot collaboration. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1-14. https://doi.org/10.1145/3411764.3445355

Manera, V., Becchio, C., Cavallo, A., Sartori, L., & Castiello, U. (2011). Cooperation or competition? Discriminating between social intentions by observing prehensile movements. *Experimental Brain Research, 211*(3-4), 547-556. https://doi.org/10.1007/s00221-011-2649-4

Marsh, K. L., Richardson, M. J., & Schmidt, R. C. (2009). Social connection through joint action and interpersonal coordination. *Topics in Cognitive Science, 1*(2), 320-339. https://doi.org/10.1111/j.1756-8765.2009.01022.x

Mattar, M. G., & Lengyel, M. (2022). Planning in the brain. *Neuron*. https://doi.org/10.1016/j.neuron.2021.12.018

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought.* University of Chicago Press. https://doi.org/10.1515/9783110874259.351

McNeil, N. M., Alibali, M. W., & Evans, J. L. (2000). The role of gesture in children's comprehension of spoken language: Now they need it, now they don't. *Journal of Nonverbal Behavior, 24*(2), 131-150. https://doi.org/10.1023/A:1006657929803

Melinger, A., & Kita, S. (2007). Conceptualisation load triggers gesture production. *Language and Cognitive Processes, 22*(4), 473-500. https://doi.org/10.1080/01690960600696916

Melis, A. P., Hare, B., & Tomasello, M. (2006). Chimpanzees recruit the best collaborators. *Science, 311*(5765), 1297-1300. https://doi.org/10.1126/science.1123007

Miller, K. J., & Venditto, S. J. C. (2021). Multi-step planning in the brain. *Current Opinion in Behavioral Sciences, 38*, 29-39. https://doi.org/10.1016/j.cobeha.2020.07.003

Nalepka, P., Kallen, R. W., Chemero, A., Saltzman, E., & Richardson, M. J. (2017). Herd those sheep: Emergent multiagent coordination and behavioral-mode switching. *Psychological Science, 28*(5), 630-650. https://doi.org/10.1177/0956797617692107

Newman-Norlund RD, van Schie HT, van Zuijlen AMJ, Bekkering H (2007) The mirror neuron system is more active during complementary compared with imitative action. *Nature Neuroscience, 10*(7), 817–818. https://doi.org/10.1038/nn1911

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science, 304*(5669), 452-454. https://doi.org/10.1126/science.1094285

Özyürek, A., Willems, R. M., Kita, S., & Hagoort, P. (2007). On-line integration of semantic information from speech and gesture: Insights from event-related brain potentials. *Journal of Cognitive Neuroscience, 19*(4), 605-616. https://doi.org/10.1162/jocn.2007.19.4.605

Pezzulo, G., & Dindo, H. (2011). What should I do next? Using shared representations to solve interaction problems. *Experimental Brain Research, 211*(3-4), 613-630. https://doi.org/10.1007/s00221-011-2712-1

Pezzulo, G., Donnarumma, F., Dindo, H., D'Ausilio, A., Konvalinka, I., & Castelfranchi, C. (2019). The body talks: Sensorimotor communication and its brain and kinematic signatures. *Physics of Life Reviews, 28*, 1-21. https://doi.org/10.1016/j.plrev.2018.06.014

Prinz, W. (1997). Perception and action planning. *European Journal of Cognitive Psychology, 9*(2), 129-154. https://doi.org/10.1080/713752551

Rauscher, F., Krauss, R. M., & Chen, Y. (1996). Gesture, speech, and lexical access: The role of lexical movements in speech production. *Psychological Science, 7*(4), 226-231. https://doi.org/10.1111/j.1467-9280.1996.tb00364.x

Richardson, M. J., Marsh, K. L., & Baron, R. M. (2007). Judging and actualizing intrapersonal and interpersonal affordances. *Journal of Experimental Psychology: Human Perception and Performance, 33*(4), 845–859. https://doi.org/10.1037/0096-1523.33.4.845

Richardson, M. J., Marsh, K. L., Isenhower, R. W., Goodman, J. R., & Schmidt, R. C. (2007). Rocking together: Dynamics of intentional and unintentional interpersonal coordination. *Human Movement Science, 26*(6), 867-891. https://doi.org/10.1016/j.humov.2007.07.002

Rizzolatti, G., & Arbib, M. A. (1998). Language within our grasp. *Trends in Neurosciences*, *21*(5), 188-194. https://doi.org/10.1016/S0166-2236(98)01260-0

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Reviews of Neuroscience, 27*, 169-192. https://doi.org/10.1146/annurev.neuro.27.070203.144230

Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research, 3*(2), 131-141. https://doi.org/10.1016/0926-6410(95)00038-0

Rizzolatti, G., & Sinigaglia, C. (2010). The functional role of the parieto-frontal mirror circuit: interpretations and misinterpretations. *Nature Reviews Neuroscience, 11*(4), 264-274. https://doi.org/10.1038/nrn2805

Roberts, M. E., & Goldstone, R. L. (2011). Adaptive group coordination and role differentiation. *PLoS One, 6*(7), e22377. https://doi.org/10.1371/journal.pone.0022377

Roepstorff, A., & Frith, C. (2004). What's at the top in the top-down control of action? Script-sharing and 'top-top' control of action in cognitive experiments. *Psychological Research, 68*(2-3), 189-198. https://doi.org/10.1007/s00426-003-0155-4

Sacheli, L. M., Tidoni, E., Pavone, E. F., Aglioti, S. M., & Candidi, M. (2013). Kinematics fingerprints of leader and follower role-taking during cooperative joint actions. *Experimental Brain Research, 226*(4), 473-486. https://doi.org/10.1007/s00221-013-3459-7

Sartori, L., Becchio, C., Bulgheroni, M., & Castiello, U. (2009). Modulation of the action control system by social intention: unexpected social requests override preplanned action. *Journal of Experimental Psychology: Human Perception and Performance, 35*(5), 1490. https://doi.org/10.1037/a0015777

Sartori, L., Becchio, C., & Castiello, U. (2011). Cues to intention: the role of movement information. *Cognition, 119*(2), 242-252. https://doi.org/10.1016/j.cognition.2011.01.014

Schmidt, R. C., Carello, C., & Turvey, M. T. (1990). Phase transitions and critical fluctuations in the visual coordination of rhythmic movements between people. *Journal of Experimental Psychology: Human Perception and Performance, 16*(2), 227. https://doi.org/10.1037/0096-1523.16.2.227

Schmidt, R. C., & O'Brien, B. (1997). Evaluating the dynamics of unintended interpersonal coordination. *Ecological Psychology, 9*(3), 189-206. https://doi.org/10.1207/s15326969eco0903_2

Schuch, S., & Tipper, S. P. (2007). On observing another person's actions: Influences of observed inhibition and errors. *Perception & Psychophysics, 69*(5), 828-837. https://doi.org/10.3758/BF03193782

Sebanz, N., Bekkering, H., & Knoblich, G. (2006). Joint action: bodies and minds moving together. *Trends in Cognitive Sciences, 10*(2), 70-76. https://doi.org/10.1016/j.tics.2005.12.009

Sebanz, N., & Knoblich, G. (2009). Prediction in joint action: What, when, and where. *Topics in Cognitive Science, 1*(2), 353-367. https://doi.org/10.1111/j.1756-8765.2009.01024.x

Sebanz, N., Knoblich, G., & Prinz, W. (2003). Representing others' actions: just like one's own?. *Cognition, 88*(3), B11-B21. https://doi.org/10.1016/S0010-0277(03)00043-X

Sebanz, N., Knoblich, G., & Prinz, W. (2005). How two share a task: corepresenting stimulus-response mappings. *Journal of Experimental Psychology: Human Perception and Performance, 31*(6), 1234. http://dx.doi.org/10.1037/0096-1523.31.6.1234

Sebanz, N., Knoblich, G., Prinz, W., & Wascher, E. (2006). Twin peaks: An ERP study of action planning and control in co-acting individuals. *Journal of Cognitive Neuroscience, 18*(5), 859–870. https://doi.org/10.1162/jocn.2006.18.5.859

Shintel, H., & Keysar, B. (2009). Less is more: A minimalist account of joint action in communication. *Topics in Cognitive Science, 1*(2), 260-273. https://doi.org/10.1111/j.1756-8765.2009.01018.x

Sievers, B., Welker, C., Hasson, U., Kleinbaum, A. M., & Wheatley, T. (2020). *How consensus-building conversation changes our minds and aligns our brains.* PsyArXiv. https://doi.org/10.31234/osf.io/562z7

Singer, M. A., & Goldin-Meadow, S. (2005). Children learn when their teacher's gestures and speech differ. *Psychological Science, 16*(2), 85-89. https://doi.org/10.1111/j.0956-7976.2005.00786.xs

Simon, J. R. (1990). The effects of an irrelevant directional cue on human information processing. *Advances in Psychology, 65*, 31-86. https://doi.org/10.1016/S0166-4115(08)61218-2

Snider, J., Lee, D., Poizner, H., & Gepshtein, S. (2015). Prospective optimization with limited resources. *PLoS Computational Biology*, *11*(9), e1004501. https://doi.org/10.1371/journal.pcbi.1004501

Tollefsen, D. (2005). Let's pretend! Children and joint action. *Philosophy of the Social Sciences, 35*(1), 75-97. https://doi.org/10.1177/0048393104271925

Tomasello, M. (2010). *Origins of human communication.* MIT Press.

Tomasello, M., & Carpenter, M. (2007). Shared intentionality. *Developmental Science, 10*(1), 121-125. https://doi.org/10.1111/j.1467-7687.2007.00573.x

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences, 28*(5), 675-691. https://doi.org/10.1017/S0140525X05000129

Uithol, S., van Rooij, I., Bekkering, H., & Haselager, P. (2011). Understanding motor resonance. *Social Neuroscience, 6*(4), 388-397. https://doi.org/10.1080/17470919.2011.559129

van der Wel, R. P., Knoblich, G., & Sebanz, N. (2011). Let the force be with us: dyads exploit haptic coupling for coordination. *Journal of Experimental Psychology: Human Perception and Performance, 37*(5), 1420–1431. https://doi.org/10.1037/a0022337

van Schie, H. T., van Waterschoot, B. M., & Bekkering, H. (2008). Understanding action beyond imitation: reversed compatibility effects of action observation in imitation and joint action. *Journal of Experimental Psychology: Human Perception and Performance, 34*(6), 1493-1500. https://doi.org/10.1037/a0011750

Vesper, C., & Richardson, M. J. (2014). Strategic communication and behavioral coupling in asymmetric joint action. *Experimental Brain Research, 232*(9), 2945-2956. https://doi.org/10.1007/s00221-014-3982-1

Wameken, F., Chen, F., & Tomasello, M. (2006). Cooperative activities in young children and chimpanzees. *Child Development, 77*(3), 640-663. https://doi.org/10.1111/j.1467-8624.2006.00895.x

Warneken, F., Hare, B., Melis, A. P., Hanus, D., & Tomasello, M. (2007). Spontaneous altruism by chimpanzees and young children. *PLoS Biology, 5*(7), e184. https://doi.org/10.1371/journal.pbio.0050184

Warneken, F., & Tomasello, M. (2007). Helping and cooperation at 14 months of age. *Infancy, 11*(3), 271-294. https://doi.org/10.1111/j.1532-7078.2007.tb00227.x

Wesp, R., Hesse, J., Keutmann, D., & Wheaton, K. (2001). Gestures maintain spatial imagery. *The American Journal of Psychology, 114*(4), 591-600. https://doi.org/10.2307/1423612

Willems, R. M., Özyürek, A., & Hagoort, P. (2007). When language meets action: The neural integration of gesture and speech. *Cerebral Cortex, 17*(10), 2322-2333. https://doi.org/10.1093/cercor/bhl141

Wilson, M., & Knoblich, G. (2005). The case for motor involvement in perceiving conspecifics. *Psychological Bulletin, 131*(3), 460-473. https://doi.org/10.1037/0033-2909.131.3.460

Yamamoto, S., Humle, T., & Tanaka, M. (2012). Chimpanzees' flexible targeted helping based on an understanding of conspecifics' goals. *Proceedings of the National Academy of Sciences, 109*(9), 3588-3592. https://doi.org/10.1073/pnas.1108517109

Zacks, J. M., Tversky, B., & Iyer, G. (2001). Perceiving, remembering, and communicating structure in events. *Journal of Experimental Psychology: General, 130*(1), 2. https://doi.org/10.1037/0096-3445.130.1.29

Appendix A

Coding Scheme of Content of Speech and Examples of Each Category

Advanced the task

Joint model of the task

Established an element of a joint action

Part and structure

Identity and property of a part or tool

- This is the bottom.
- The holes switch in the middle.
- So this goes this side.
- The holes sit at the bottom.
- That fits here.

State of the structure

- It does stay in.
- Good here.
- Only one side in.
- (This is in?) I think so, yeah.
- Doesn't look like it's all the way in.

Action

Who does what

- So let me put the top board first.
- You hold this.
- You do the screw and I'll try to hold it to get out.
- I squeeze a little bit so I can screw in.
- You want to hold it and I can put it in?

What to do

- We'll put the wheels on last.
- Let's get the main framework first.
- Shall we put this in?
- I think we should turn this a little.
- Can we just put the top one?

Phase

- We'll put the wheels on last.
- Let's get the main framework first.
- You want to put all of these, all the side pieces on one of these and the other one on the other side?

Step

- Shall we put this in?
- Can we just put the top one?
- Let's do the bottom one.

- So we can screw this on the other side.
- Put the wheels.

Sub-action

- You hold this.
- You do the screw and I'll try to hold it to get out.
- I think we should turn this a little.
- I squeeze a little bit so I can screw in.
- You want to hold it and I can put it in?

Incomplete clause

- This is I guess [action of putting a side board in place].
- They are like [action of putting a side board along the correct direction].
- Well, I actually just put the [action of putting the bottom board in place].
- So that's [pointing to the screws and their correct locations on the boards].
- This thing goes [action of putting the support board in place].

Clarification for confusion

- This is in?
- Is it these things, or is it…?
- Or maybe we can turn it around.

- So which side shall we stack first?
- There's only one screwdriver?

Acknowledgment

- Okay.
- Yeah.
- Sure.
- It's okay, thanks.
- That would be good.

Did not advance the task

- Because this seems [without any gesture or action].
- Stand up.
- Check that out.
- I don't know.
- Let's see.

Appendix B

Coding Scheme of Joint Action

Joint action

- Major complementary action: Time periods during which both participants worked on attaching the same parts in an interdependent manner to achieve a common local goal.
    - *Top-side*: Time period during which the two participants assembled the top shelf and a side board.
    - *Support-side*: Time period during which the two participants assembled the support board and a side board.
    - *Bottom-side*: Time period during which the two participants assembled the bottom shelf and a side board.
    - *Wheel-structure*: Time period during which the two participants assembled the wheels and the structure-so-far.
- Sub-action: Joint behavioral primitive comprising a major complementary action.
    - *Assemble-assemble*: Time periods during which both participants acted to put the parts together.
    - *Assemble-hold*: Time periods during which participant Left acted to put the parts together while participant Right held the parts and structure-so-far in place.
    - *Hold-assemble*: Time periods during which participant Right acted to put the parts together while participant Left held the parts and structure-so-far in place.
    - *Hold-hold*: Time periods during which both participants held the parts and structure-so-far in place.

Direction of assembling for each joint action

- *Doing*: Joint actions serving to add more parts to the assembled structure.
- *Undoing*: Joint actions serving to disassemble the structure-so-far.

Coordination of each joint action

- *Speech*: When a joint action was coordinated by preceding utterances only.
    - *Who* does what: The preceding utterances specified division of labor between the participants.
    - *What* to do: The preceding utterances established the local common goal for the participants.
- *Gesture*: When a joint action was coordinated by preceding gestures only.
- *Both speech and gesture*: When both preceding utterances and gestures served to coordinate a joint action.
    - *Who* does what: The preceding utterances specified division of labor between the participants.
    - *What* to do: The preceding utterances established the local common goal for the participants.
- *Implicitly*: When a joint action was coordinated by neither preceding utterances nor gestures.

Appendix C

Table 3: Interrater reliability

	Codes	Kappa	Percent agreement
Content of Speech	Advanced Did not advance	.71	.94
	Joint model Acknowledgment	.84	.93
	Established Clarified	.84	.94
	Part and structure Action Incomplete	.82	.91
	Identity and property State of structure	.87	.95
	Who What	.79	.9
Gesture	Deictic Representational Head movement	.81	.99
	Overlapping Independent	.88	.96
	Complementary Redundant	.85	.88
Individual action	Top-side Support-side Bottom-side Wheel-structure	.92	.94
Major action	Complementary Supplementary	.82	.93
Sub-action	Assemble-assemble Assemble-hold Hold-assemble Hold-hold	.86	0.90
Direction of assembly	Doing Undoing	.93	.99
Coordination of major action	Speech-what Speech-who Gesture Implicitly	.84	.90
Coordination of sub-action	Speech-what Speech-who Gesture Implicitly	.97	.99

Appendix D

Unreported Data and Results

In addition to the reported coding scheme of content of speech, we coded 1) how each established elements was acknowledged, that is, whether or not it was acknowledged by explicit means (kappa = .869, $p < .001$, percentage agreement = .933) and whether each of them caused any change in the listener's action (kappa = .810, $p < .001$, percentage agreement = .918), 2) the scale of the action that each action-element-establishing utterance coordinated, that is, sub-action (behavioral primitive), step (major action), and phase (two or more major actions) (kappa = .787, $p < .001$, percentage agreement = .828), and 3) the fate of each action-element-establishing utterance, that is, whether an utterance ended up guiding a subsequent action, overridden at a later time, or serving as explanation for a preceding or the current action (kappa = .941, $p < .001$, percentage agreement = 972). We also compared between the rates of utterances occurring different quarters of the task.

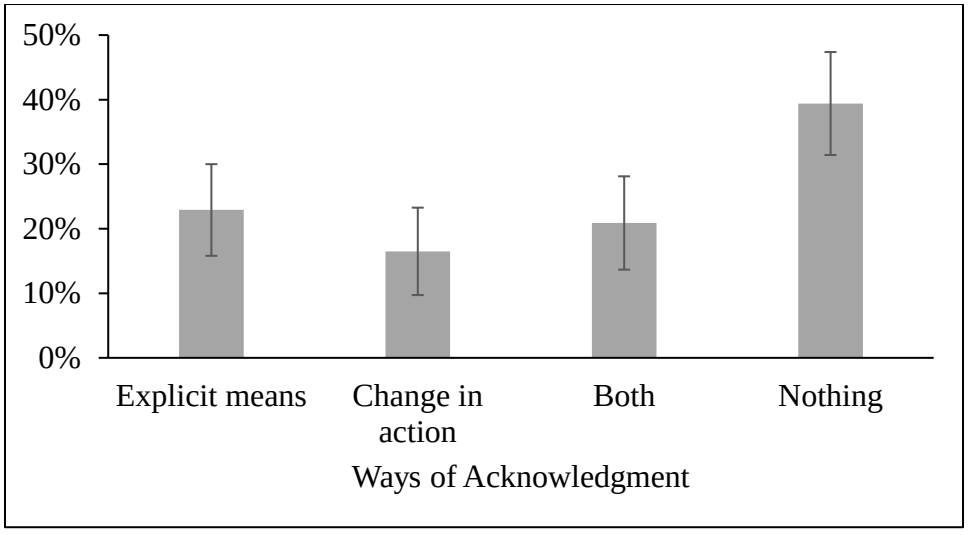

Figure 15: Ways the established elements were acknowledged.

As in Figure 15, the rate of established elements acknowledged by nothing was higher than those of established elements established by explicit means ($F(1, 12) = 7.90$, $p = .016$, η^2

= .40) and change in action ($F(1, 12) = 10.70$, $p = .007$, $\eta^2 = .47$). The rate of established elements acknowledged by nothing was boundary higher (after adjusting the alpha value to .0167) than that of established elements acknowledged by both explicit means and change in action ($F(1, 12) = 6.437$, $p = .026$, $\eta^2 = .35$). These results suggested the co-actors tended to interpret an element of a joint action as grounded and shared by both parties if no objection or question was raised.

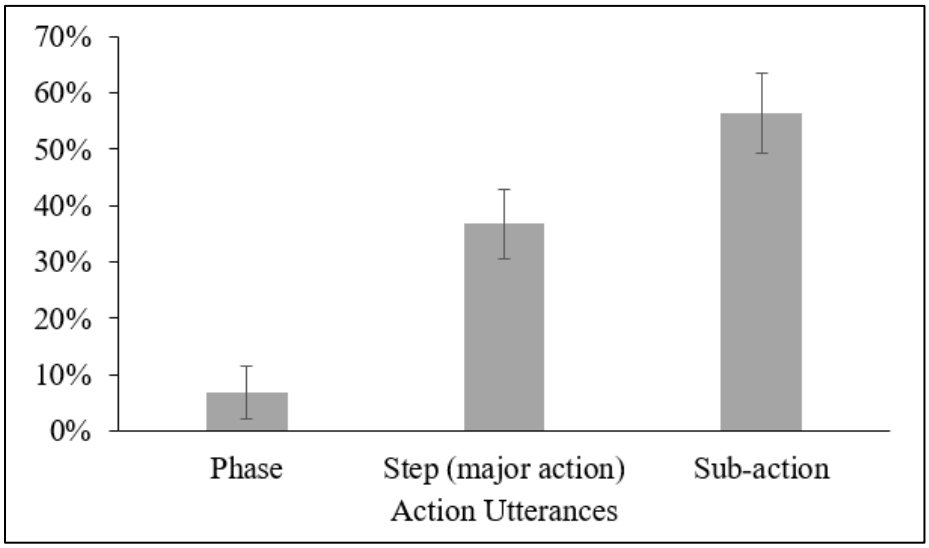

Figure 16: Scale of actions coordinated by action-element-establishing utterances.

As in Figure 16, the rate of action utterances differed across different scales of actions that they coordinated ($F(2, 24) = 36.45$, $p < .001$, $\eta^2 = .75$). The rate of action utterances that coordinated a sub-action was higher than the rates of action utterances that coordinated a phase of the task ($F(1, 12) = 74.39$, $p < .001$, $\eta^2 = .86$) and a step ($F(1, 12) = 8.00$, $p = .015$, $\eta^2 = .40$).

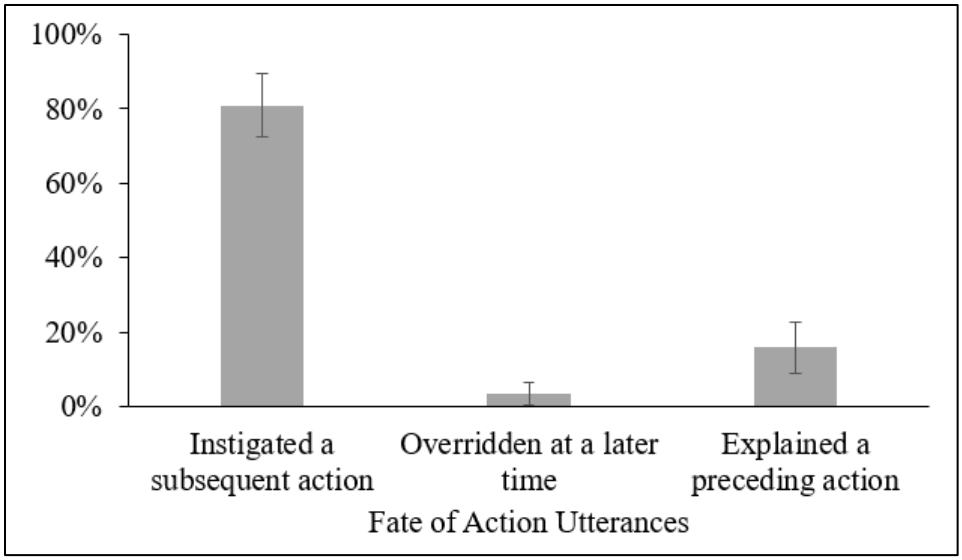

Figure 17: Fate of action-element-establishing utterances.

As in Figure 17, the rate of action utterances that ended up instigating a subsequent joint action was higher than those of action utterances ending up overridden ($F(1, 12) = 168.26$, $p < .001$, $\eta^2 = .93$) and serving as explanation ($F(1, 12) = 59.77$, $p < .001$, $\eta^2 = .83$).

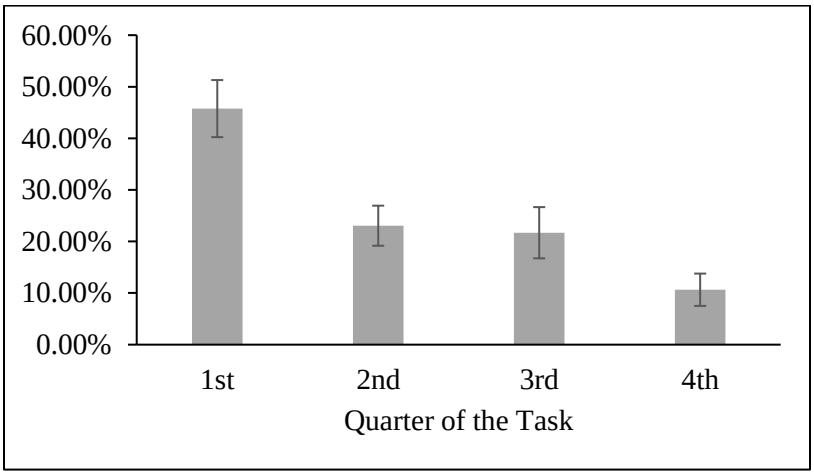

Figure 18: Rate of utterances across the four quarters of the task.

As in Figure 18, the rate of speech declined over the four quarters of the task ($F(3, 36) = 24.52$, $p < .001$, $\eta^2 = .67$). The rate of utterances occurring in the first quarter was higher than that of utterances occurring in the second quarter ($F(1, 12) = 32.84$, $p < .001$, $\eta^2 = .73$); the difference between the second and the third quarter was nonsignificant ($F(1, 12) = .11$, $p = .75$, $\eta^2 = .009$);

and the rate of utterances occurring in the third quarter was higher than that of utterances occurring in the fourth quarter (M = 10.64%, SD = 6.53%; $F(1, 12) = 12.89$, $p = .004$, $\eta^2 = .52$).