# 8

# Natural Forms of Purposeful Interaction among Humans

## What Makes Interaction Effective?

Stephen C. Levinson

**Abstract**

The design of an interactive robot should make crucial reference to the observed properties of human interaction. Obviously, human communicative interaction varies across languages and cultures, but remarkably uniform is the basic organization of interactive language use: participants take short turns at talking while avoiding overlap; they utilize a basic inventory of action–response pairs (e.g., question–answer), which can be recursively employed; they have systematic backup systems for communicative difficulties and deploy multimodal signals (speech, gesture, facial expression, gaze) to disambiguate or reinforce intended content. This chapter spells out these design properties and makes the point that human comprehension is fundamentally predictive, and has to be so to achieve the typically rapid response times despite the large latencies involved in generating speech. These properties may pose a substantial, even insuperable, hurdle for a fully humanoid interactive robot, but fortunately humans are excellent at adapting to interactants with restricted capabilities, such as children, foreigners, or aphasics.

### Introduction

Humans appear to involve themselves effortlessly in social interaction with their peers and are capable of rapidly integrating novel tasks and routines into these interactions. Clearly, efforts to build machines that might assist humans in varied tasks have much to learn from an analytic grasp of what makes humans so seamlessly able to conduct cooperative task management, even when it is novel. Right at the start, a caveat is in order: the domain of the study of human communicative interaction is still in its infancy; it is a field that has been dominated by a few maverick pioneers, and has only recently

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

acquired the extensive public databases and measurements typical of cumulative science. Although our understanding of human–human interaction is still quite limited, what we know already makes the prospects for seamless human–machine interaction quite remote. I will at the end, however, suggest that emulating humans may not in fact be the most productive use of new technologies, if indeed it is possible at all. The body of this chapter tries to delineate what we know about human interactive skills (see also Thomaz et al., this volume).

The proverbial Martian arriving on Earth would quickly notice that humans have the propensity to huddle in a face-to-face arrangement and engage in a curious exchange of alternating short bursts of communicative activity. S/he would also note that in this regard there are plenty of parallels with other species (e.g., the vocal duetting of many types of songbirds and many species of primates). Those other species tend, however, to have a small, relatively fixed repertoire of signals, whereas it would rapidly become self-evident that such is not the case for humans. Human repertoires are not only immense, they also vary significantly across ethnic groups or cultures. Moreover, it would be obvious that human exchanges happen in myriad different contexts, apparently aiding numerous types of endeavor. Beyond that, the system might be inscrutable. Let us try here to analytically unpack this a bit.

## Basic Ethological Properties of Human Communicative Interaction

The fundamental niche for human communication is social interaction in a face-to-face context: this is the context in which language is learned, the bulk of usage occurs, and almost certainly the context in which it has evolved. It is characterized by the rapid exchange of alternating short bursts of communication (averaging ca. 2 sec) as well as by multimodality: the face, the hands, the deployment of the trunk as well as the vocal organs are typically all in play at once. One can look at the system from the point of view of comprehension, in which case it is clear that the incoming multimodal signal is parsed in parallel and integrated extremely fast, also combining with numerous aspects of the context. Gestures, for example, can be shown to be unified with the linguistic message just as fast as they happen (Özyürek et al. 2007). From a production perspective, the language system may be more serial (Indefrey 2011): a message is composed and serially encoded, although the processing of each successive chunk can proceed in parallel as it passes through the many stages of encoding from message to linguistic form to articulation. However, once multimodal production is considered, it is clear that facial expressions, manual gestures, and other bodily components must be produced in parallel but temporally integrated, much like a chamber orchestra would perform a score.

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

Although the strictly linguistic aspects of comprehension have been extensively studied experimentally, the multimodal aspects remain fairly obscure. Linguistic "strings" are often treated as linear—phonemes follow phonemes, words follow words—but clearly multimodal signals are delivered simultaneously and, in this regard, are like prosody and voice quality, offering different kinds of units at different parallel levels, but somehow integrated semantically and temporally. In addition, comprehension and production must work closely in consort, as the following considerations show, and this area has only recently started to be explored.

As mentioned, interactive communication involves the rapid alternation of speaking roles. What is interesting about this is the cognitive load that is involved. Across languages and across conversational corpora, the modal gap between turns is only 100–200 msec, quite literally in the blink of an eye. It takes at least 600 msec to crank up the language production machinery; that is, the time it takes from knowing what word you want to say until the time anything comes out of your mouth. For a simple clause, the latency is more like 1500 msec (for references, see Levinson 2016). The implication is clear: to respond so rapidly, the speaker must predict the content of the incoming turn and start early preparation of a response, as illustrated schematically in Figure 8.1. We have shown that the point of an utterance or the speech act is often predicted from the very first words of an utterance (Gisladottir et al. 2015). We have also shown, by using EEG, that the production system starts as soon as the point of the incoming turn becomes clear, as indicated in Figure 8.1 (Bögels et al. 2015b) and then proceeds all the way through the various encoding stages. Thus, predictive comprehension and language production have to work in overlap and in consort. Humans are not generally good at multitasking,[1] due in part to the working memory bottleneck, so turn-taking must impose a heavy cognitive load.

The structure of interaction involves sequences of speech acts (i.e., actions packaged up in linguistic and multimodal format). Humans clearly map linguistic utterances into something action-like. Consider, for instance, "Can you reach the wine?" and its nonverbal response action. The pragmatic thrust, the point of an utterance, is only very indirectly related to its form. Thus English yes–no questions typically come in declarative format with falling prosody; the giveaway is often the manner in which the declarative is a statement about something that is more within the addressee's epistemic domain (e.g., "you are feeling better"). The many-to-many mapping between linguistic form and speech act has been explored in corpora (e.g., Couper-Kuhlen 2014; Levinson 2013a, 2017), and although inference can make use of varying probabilities, it is abductive in character involving many contextual parameters. For example,

[1] It is generally agreed that multitasking slows performance and increases errors, but the idea that true multitasking is impossible has required recent revision; for a review, see Fischer and Plessow (2015).

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

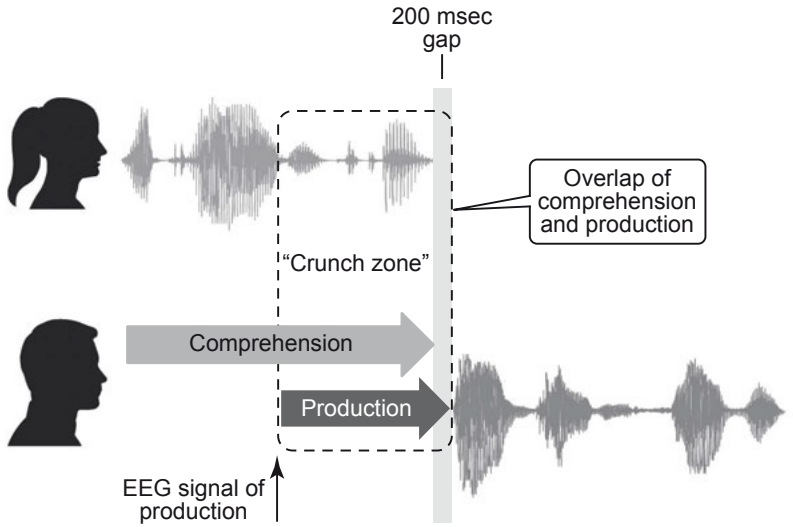


**Figure 8.1** Overlapping comprehension and production processes during turn-taking (after Levinson 2016).

an utterance like "Can we lift this?" could be a request for permission to remove something, an offer to help you with your suitcase, or an enquiry about the weight of something; apart from prosody, only contextual factors are likely to disambiguate. Outside restricted domains (e.g., travel agents, bank enquiries) these inferences will be very challenging to model in human–machine interaction.[2]

As noted, speech actions come in sequences and often occur as two paired actions (a so-called adjacency pair): following a question, an answer is due; following an offer, an acceptance or refusal is expected; following a request, an action or excuse will be forthcoming (see Exchange 1 and 2 below). Actions are thus often contingent on prior actions: "yes" only makes sense in relation to the prior query. The simple device of paired actions can, however, be recursively applied according to the template shown in Figure 8.2 (Kendrick et al., in prep.): FPP marks the first part of a paired sequence of actions and SPP marks the second part, the response; each of the expansion types can also consist of pairs of actions (Schegloff 2007).

Consider the following exchange involving a question–answer pair embedded within a question–answer pair:

[2] A reviewer made the interesting point that the complexity of inference is narrowed in humans by the matching evolved design of sensors and effectors, and no doubt the cognition connecting them: I know what is visually salient to you without complex calculation, but this may not be available to machines with their different perceptual systems. This may be a more serious barrier than inference to constructing satisfactory interacting machines.

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

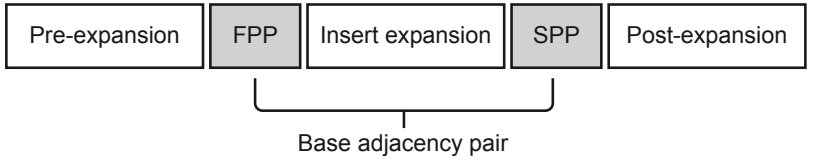


**Figure 8.2** Basic template for sequence organization. FPP marks the first part of a paired sequence of actions, and SPP the second part, the paired response. Each expansion type can also consist of pairs of actions. After Schegloff (2007).

Exchange 1
A: "May I have a bottle of Mich?"
B: "Are you twenty one?"
A: "No"
B: "No"

Interestingly this kind of center embedding can go on indefinitely—naturally occurring cases of at least six degrees of center embedding have been found in conversation, far more complex than anything found in natural language syntax where such embeddings are capped at level 3, and then only in written language (see Levinson 2013b). A lot of ink has been spilt on why the limitation is found in syntax, and it is usually attributed to short-term memory problems. What is interesting, then, is that these limitations do not hold in a joint task, even though each party must hold in mind the same pushdown stack model. This is a first indication that there is something rather special about the human capacity for joint undertakings, something that seems to be largely absent from other species: it is a capacity to "distribute cognition" over individuals (Hollan et al. 2000; Hutchins 1995), thus allowing interactants to form a joint computational device, which seems able to overcome the stack-maintaining limitations of individuals.

Center embedding is also exploited for a fundamentally important function; namely, interactive repair. This typically involves a repair initiator (e.g., "huh"? or "what"?) from the recipient followed by a repeat or clarification of what has just been said by the first speaker. Repair becomes much more difficult if displaced from the slot immediately after the troublesome turn. Thus there is pressure to solve understanding or hearing problems as soon as possible. Without this basic repair mechanism, our understandings would rapidly diverge; repair, therefore, plays a crucial function in maintaining intersubjectivity or mutual understanding, occurring roughly every 80 sec in any language (see Dingemanse et al. 2015). This kind of repair uses the "insert expansion" slot in Figure 8.2 and can be recursive, so this can get quite complex, as in the following center-embedded insert expansions (Merritt 1976; see also Levinson

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

2013b), where the pushdown stack character is indicated by the indentation and line numbering:[3]

Exchange 2

S: Next ← *Request to order*
0 C: Roast beef on rye ← *Order*

1 S: Mustard or mayonnaise? ← *Q1*
2 C: Excuse me? ←*Repair–Initiator* (*R1*)
3 S: What? ← *Repair–Initator* (*R2*) *on R1*

3 C: Excuse me? ←*Repair of R2*
2 I didn't hear what you said
1 S: Do you want mustard or mayonnaise? ← Repair of Q1, requested in R1
C: Mustard please. ← *A1 to Q1*

0 S: ((provides)) ← *Compliance with order in line 0*

Other kinds of sequence exist. For example, one possibility is to exploit the preexpansion slot in the schema above (Figure 8.2), as in Exchange 3 (below), where a prior adjacency pair checks the preconditions for the main action (in this case, an offer). Here, the pre-FPP ("Say, whadja doing?") is plausibly a pre-offer, but it could also be a pre-request (a preliminary to, e.g., "Want to come and help me clean up?"): B's response is not a straight answer (compared to "doing homework" or the like), but is couched to encourage the following action, which has clearly been foreseen. Sequences thus involve a kind of look ahead, with responses geared to inferences about what is likely to be coming up next. From the perspective of artificial intelligence (AI), this can be thought of as plan reconstruction: B infers that A has a plan to offer something nice, and so encourages it; likewise, A projects from B's response that B intends to do nothing to impede the offer and is therefore likely to accept it. This meshing of mutually inferable plans is a crucial property of human interaction, both verbal and nonverbal (e.g., constructing something together) (Bangerter and Clark 2003).

Exchange 3

1 A: "Say, whadja doing?"
2 B: "Not much"
3 A: "Y'wanna drink?"
4 B: "Yeah"

[3] One reviewer queried the recursive structure here, suggesting a possible list representation; see Levinson (2013b) for many examples, which I think have to be understood as interactively checking preconditions on preconditions on preconditions… (up to 6 deep) to actions. A list structure cannot capture the way in which each "push" has to have its own paired "pop," as in the structure $[A_1\ [B_1\ [C_{1\ldots}C_2]\ B_2]\ A_2]$; see Levinson (2013b).

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

How exactly these plans are made mutually manifest, and how far downstream interlocutors predict the likely courses of action, is a puzzle (see Levinson 2013a, b, 2017). Clearly, as in the example above, sequences may have a recurrent pattern, and just as observers of primate behavior have noted under the rubric of *ontogenetic ritualization*, the initial action can come to project the entire sequence. In human interaction, however, such projection is often much less clear, and as the potential ambiguity of the pre-offer/pre-request above indicates, the stakes for misconstruals can be relatively high.

In work we have conducted across languages and cultures, all of these patterns (turn-taking, repair, sequence organization) are strongly universal, following the same principles in many detailed ways (e.g., Dingemanse et al. 2015; Kendrick et al., in prep.; Stivers et al. 2009). This contrasts markedly with the diversity of languages, which differ at every level, from the sound system, to the combinatorics involved in phonology, morphology and syntax, to the meanings that are conveyed. We believe these interactional principles form a strong infrastructure for language, which in turn makes the learning of languages possible and influences their structure in subtle ways. In addition, all spoken languages exhibit deep similarities in the use of multimodal resources despite the occasional cultural taboos to be found here: gaze, gesture, and facial expressions play an important role in framing and supplementing the linguistic content.

The properties of human communicative interaction reviewed here can be abstracted into a set of design features (see Table 8.1). These are some of the basic desiderata that any interactive computational agent will either have to mimic or be endowed with ways to achieve equivalent functionality. Table 8.1 sketches the functions of each design feature, which we briefly review.

Multimodality offers redundancy or complementarity in, for instance, gesture, facial expression, and speech. Crucially, the taking of short turns will make clear, in your immediate response, whether you understood me correctly ("legibility"). Action sequences, such as question–answer (Q–A), structure interaction by setting up expectations for responses; note that in Exchange 1, the expectation can be postponed but still persists. Although spoken turns minimize overlap, nonvocal signals (like laughing, smiling, shaking hands) may get their significance precisely through simultaneity. Communication about the state of communication (metacommunication, as in repair initiators) plays a crucial cybernetic role in guiding utterance interpretation and signaling communicative success. The fact that conversation has an expected "clock speed" allows participants to sense an interpretative problem simply from a delayed response; the expected simultaneity of, for instance, laughter is a further check on temporal meshing. Feedback signals like "mm-hmm" or nodding typically occur in overlap at the end of an utterance constituent, and their timing allows the speaker to proceed rapidly. None of this would work without a presumption of engagement and engagability, which makes it possible to enter an interaction with a stranger on the street. This presumptive mutual regard and helpfulness

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

**Table 8.1** Key design features of human interaction: a checklist for constructing intelligent interactive agents.

| | |
|---|---|
| 1. Media | |
| • Function: communication, redundancy | Multimodal signals, language |
| 2. Action sequences | |
| a. Alternating turns | Speech act mapped to language |
| • Function: includes legibility, opportunities for repair | |
| b. Action sequences | Adjacency pairs (e.g., Q–A) |
| • Function: includes structuring exchange | Complex sequences, e.g., insert pairs (e.g., Q–Q–A–A) |
| c. Simultaneous<br>• Function: coordination, ritual | Complex sequences, e.g., shaking hands, laughing |
| 3. Metacommunication | |
| • Function: check communication | Repair |
| • Function: confirm message receipt | Feedback tokens, e.g., *uhuh* |
| 4. Timing | |
| • Function: indicates state of processing is "on time" | Turn-taking timing |
| • Function: "clock speed" check | Synchronicity, e.g., shaking hands |
| • Function: "message received now" | Timing of feedback |
| 5. Motivation<br>• Function: enter and maintain engagement | Shared goals<br>Affect/attachment, politeness, specific rewards |
| 6. Legibility of | |
| • Function in attention: indicates current focus of processing | E.g., gaze readability |
| • Function in intention: aid predictive processing | E.g., gesture signal vs. instrumental action |

seems unique to humans, but of course is cemented when there are proximate shared goals or rewards. In general, for an interaction to work, each agent's actions must be performed in such a way to make them legible (i.e., interpretable) as an action intended to be perspicuous for its purpose. This is obvious for communicative signals, but holds for other kinds of joint actions, so that when, for example, you and I carry a table together, the direction of my gaze can indicate the intended direction of motion. It follows that instrumental actions which are not part of the joint endeavor (e.g., scratching one's head or coughing) should also be clearly legible as irrelevant for the joint purpose in hand.

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

## Affect, Empathy, and the Human Person

A robotic assistant can clearly be of great utility without having any deep understanding of humans; after all, humans will readily adapt to its limitations. But any machine that wishes to "pass" as capable of humanlike interactive task learning will have to know a great deal, not only about human communication but also about human nature. Human interaction is, in fact, replete with "ritual" aspects which no successful interactant can ignore. A century ago, the sociologist Durkheim (1912) suggested that religious beliefs personify society, the collective consciousness of consciousness, as he put it, so that individuals come to have as social persons a kind of sacred quality. Goffman (1959) built on this in his analysis of "interaction ritual," noting how we often treat persons with elaborate care: we pretend not to notice others' slips, belittle ourselves (e.g., walking into a lecture late, stooped over), and worry about our perceived social competence to ensure that mutual dignity or "face" is maintained both by self and other on each other's behalf. Brown and Levinson (1987) elaborated this account in a theory of "politeness" in which individuals' ritual or "face" requirements could be maintained in two rather different ways: by claiming empathy and fellowship (Durkheim's "positive rites") or by giving the other maximum *Lebensraum* (Durkheim's negative rites or avoidance rituals). Which kind of ritual is deployed depends on social closeness versus social distance (vertical or horizontal), together with some measure of the weightiness of the action or imposition. This translates directly into the choice of linguistic expressions: if I want to borrow the pen of my neighbor in a plane to fill out a landing form, I might say "Hey, I need your pen for a moment" if he's a friend, but "Excuse me, could I possibly borrow your pen just to fill in this form?" if he's a stranger. Although the whole business is wrapped up in culture-specific conventions (e.g., the bowing and honorifics of Japanese), there does seem to be a universal basis to these mini-rituals of the person. Further corpus work suggests that in choosing linguistic expressions we make quite elaborate computations of rights and duties, epistemic territories or domains of expertise, and estimations of effort or contingency (Drew and Couper-Kuhlen 2014). We do this because the failure to recognize the other's right to self-esteem is cause for offence; thus, one works with what Goffman (1959) called "the virtual offence" (the worst construal of what one is doing) and tries to stop it from happening. When it is necessary to invade another's domain, as in medical examinations, elaborate circumspection is required.

An aspect of recognizing the other as a "sacred being" is the recognition of the need for empathy among close associates. Failure to greet a person or omitting to extend condolences or congratulations is also cause for offence, as is the failure to laugh at people's jokes, appreciate their stories, or empathize with their travails. Interacting successfully with a child may involve entering, for a while, into its momentary make-believe world. In general, any interactional success here will be achieved by keeping tabs on the life courses of

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

all significant others. Interestingly, it is also crucial to keep track of common ground—the things we have told each other—requiring a record of informational exchange (complete with the reference forms used) for each such person. Failure to do so will have you classified as the party bore or the senile kinsman who repeats information already imparted. All this is self-evident to us, but will likely be opaque to an interactional machine: it will not be feasible to build in all the sensitivities and particularities of the social world that parameterize underlying social principles. Perhaps some future machines will be able to learn some of these mores, although that, in itself, poses formidable inferential problems of the kind explored in studies of child development. Still, modeling the stiff politeness of anonymous service staff may be within our grasp and a likely prerequisite for a successful interactive robot.

## Learning Interactively

Understanding how new tasks can be learned in and through interaction is the focus of this volume. From experiments on joint action, it seems that cooperative interaction relies on each participant modeling the other's unfolding action plans, transposing themselves into the other's footsteps as it were, and so co-representing the joint action. This seems to be so even if my half of the joint task has an independent timing and function (Sebanz et al. 2003, 2006; Vesper et al. 2016).

There have been interesting reports from cross-cultural studies of societies where children, for example, acquire most adult tasks simply through observation, not instruction or demonstration (Gaskins 1999; Rogoff et al. 2003). So how important communication, or indeed interaction, is in learning new tasks is perhaps not clear. In an interesting preliminary study, Laland and associates tested the learning of flint-tool knapping under different conditions: reverse engineering (by inspection of the tool), emulation (by watching production), restricted gestural communication, and vocal communication (Morgan et al. 2015). They found that the task was easily learned only when there was full communication, because certain critical tricks are not easily extracted through direct observation (e.g., in this flint-knapping case, preparing a striking platform under 90 degrees). More studies of this kind are needed, with different types of tasks, but generally this work suggests that directed communication may often be essential to the learning of skills.

Here I would like to make the point that any kind of sustained cooperative interaction presupposes communication, even if it is subliminal. For instance, maintaining synchrony in a chamber orchestra involves visual cues, as in the exaggerated, but precise, lifting of a bow to indicate an impending entrance ("now is when we begin, at this tempo"); carrying a table together involves some low-level signaling of direction ("go left now"), evidenced by an exaggerated tilt of the head to the left. These signals work by having a shape that

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

is not purely instrumental: the exaggeration of bow movement beyond what is needed for sound production or the tipping of the table greater than is needed to get around the corner. The detection of the not-just-instrumental quality of a basically instrumental action can be very subtle but allows our species to collaborate in a coordinated fashion (see Bangerter and Clark 2003; Grice 1975 on the Maxim of Manner). More generally, if there is a way to indicate "this is the way to do it" by the manner of doing, it may play a crucial role in human cultural transmission. This kind of signaling has been turned into a theory of "natural pedagogy" (Csibra and Gergely 2009), and the distinction between instrumental and noninstrumental action has been held to be central to the learning of culture, where both causally transparent and causally opaque actions need to be learned; unlike instrumental activities, table manners, for instance, should be followed, not innovated or improved upon (Clegg and Legare 2016).

## The Cognition behind the Ethology: The Gulf between AI and Human Interaction

It is not easy to reconstruct the underlying cognition that makes human interaction possible. Apparently, effortless coordination requires complex inference: if we are putting together an IKEA bookshelf, and I put the screwdriver down with the handle toward you, I may be signalling what you need next (Bangerter and Clark 2003). Such wordless communication is challenging to model. The philosophical reconstruction by Grice (1975) remains the best general attack we have: the signaler intends to cause an effect in the mind of the recipient just by getting the recipient to recognize that intention. But how does this happen? The general answer seems to be: by an inflection of manner that suggests the action was not purely instrumental. Why take the extra effort to turn around the screwdriver? Alternatively, take the case of a student who arrives late to class with a cappuccino moustache: a fellow student might vigorously wipe her own lip while gazing across at the new arrival, such that the latecomer wonders, "Why is she doing that?" This may lead to the realization that there is something anomalous about the latecomer's own lip. This seems to work by an implicit comparison to the simplest instrumental version of the action: any excessive elaboration of manner suggests communicative intent. This kind of incidental communication lies behind nearly all our coordination: I wait to walk across the road at a junction until I have caught the driver's eye, so that he and I both know we are aware of each other, and it's safe to cross the road (building on the assumption of our common minds and bodies, in a way that is inherently problematic for a robot).

The Gricean analysis involves reflexive ratiocination: I plan my signal with its manner inflection thinking that you will reconstruct the communicative intention behind it, realizing it is a signal and not (or not only) an instrumental action (thanks to the manner inflection). Many psychologists have

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

been doubtful of any such thing, pointing instead to our tendency for lazy egocentricity. To explore this, we designed the following task (de Ruiter et al. 2010): two players took turns to signal where the other should place a gaming piece on a checkerboard. They were denied all means of direct communication and could only indicate by the manner of moving their own piece where the other player should place hers. Indeed, one of them was in an fMRI scanner. Participants could solve the task, even though each trial had different properties that denied them the possibility of forming implicit conventions: they did so by choosing a route for their own piece that went over the square where the other should place her piece, and wiggling or rotating or otherwise by manner inflection indicating the solution. What we found is that the mentalizing brain areas activated by the signaler were reflected in the areas activated in the recipient, prima facie evidence that we do indeed attempt to mirror the other person's reasoning (de Ruiter et al. 2010; Noordzij et al. 2009). It would seem to work by an abductive leap from the mannered signal to the sender's likely intention, but an adequate computational model eludes us. Blokpoel (2015) proves formally that the inferences required are intractable (NP-hard) unless the set of goals and signals are highly constrained (see also Van Rooij et al. 2011). Attempts to model the interpretation of nonce or "one-off" signals by analogy also prove problematic (Blokpoel 2015).

These results hold just as much for linguistic communication as they do for nonce, nonconventional communication. That is because the *point* or *speech act* of an utterance is rarely explicit: the relations between linguistic form and speech acts are many-to-many (Levinson 2017). Consider "Say, whadja doing?" in Exchange 3 (above), which has the superficial form of a question, but its purpose is to pre-adumbrate an invitation or suggestion, a purpose perspicuous to the answerer: "Nothing much." The reconstruction of its intended point requires a projection of an upcoming sequence (offer and acceptance); that is, it involves plan reconstruction and plan-meshing, as illustrated in Figure 8.3. Here, from the utterance "Whadja doing?" Clara detects a possible plan to invite her out (the pre-invitation reading of the utterance), so she answers in such a way as to make clear that she has no impediments and is likely to accept (the go-ahead reading of her response, "Not much"). This kind of plan reconstruction was explored in early AI but has proved computationally tractable only when the alternatives were highly constrained (Allen and Perrault 1980), as Blokpoel's (2015) theoretical work predicts.

At present I believe we simply do not have adequate computational tools to model the mysteries of human communication, which fall in the domain of "the inference to the best explanation" or inspired abduction under an umbrella of reflexive reasoning. This is the kind of reasoning successfully employed in Schelling games of pure coordination, where we both get a prize if we can think of the same number without communicating (Schelling 1960).

We come now to the computational tractability of the ethological properties, like multimodality and turn-taking discussed above. Turn-taking with its

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

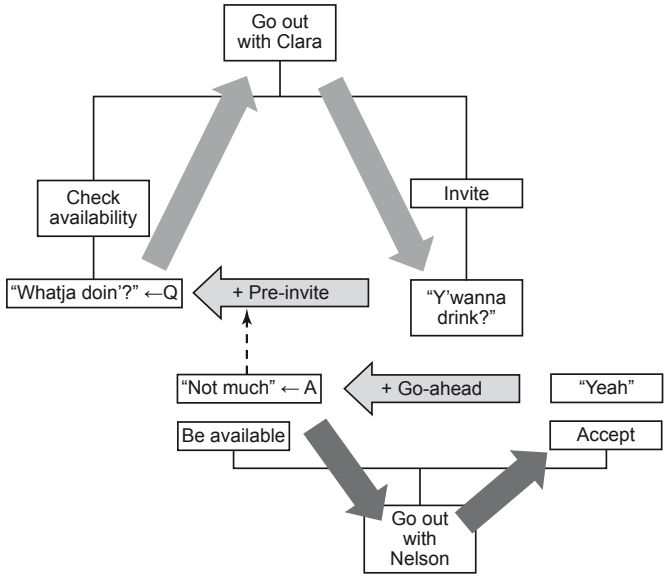


**Figure 8.3** Sequence organization as plan reconstruction (from Levinson 2017).

split-second turnaround poses problems of sheer processing power. To respond, I argue, you must compute (most likely predict) the intended point or speech act of the incoming utterance early enough to begin your response preparation well before the end of the incoming turn. This problem is fraught with the many-to-many mapping between utterance and action mentioned above. Again, we can show the inference is often made very early using brain imaging (Gisladottir et al. 2015). In the case of responses to prior utterances, decisions about the speech act may be made by the first word. Comprehension is a massively parallel undertaking, which helps to explain its speed. In addition, the pace is pressured by the fact that delay has its own semiotics: a delay over 500 msec after a request, for example, will be interpreted as reluctance to comply; using brain imaging, one can track the hearer's moment-by-moment change of expectations (Bögels et al. 2015a).

In the end, linguistic production has to be serial due to the articulation bottleneck (Indefrey 2011), as reflected by the relatively huge latencies involved in linguistic encoding. The decision about how to respond, which words and sentence frames to employ, would counterfactually predict very slow responses (consider the psychological generalization known as Hick's Law, which holds that response times increase logarithmically with the number of possible responses from which to choose, and the 30,000 odd items in the average vocabulary). So, even though production latencies are large (up to 1500+ msec to code a simple clause), given the vast decision trees possible, they are still remarkable.

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

Multimodality compounds these problems, because one is dealing with the orchestration of a veritable ensemble of channels. For instance, recent work shows that

- long blinks toward the end of an utterance signal "go on" (Hömke et al. 2017),
- gesture holds signal "turn unfinished" (Torreira and Valtersson 2015),
- gaze aversion by responder signals "this is not the response you were hoping for" (Kendrick and Holler 2017), and
- frowns on a recipient or her head thrust and freeze of body position signal "repair request coming" (Floyd et al. 2016; Kendrick 2015).

These signals often overlap with the verbal signal, and the main clue to their scope is probably timing. As in the McGurk effect,[4] it is likely that such timing can be delayed by up to 200 msec without disrupting a sense of synchronization. While massive parallel computation may again solve the comprehension problem, the orchestrated production of these signals by an early utterance plan is not understood at all (for a discussion on gesture planning, see de Ruiter et al. 2012).

All this takes place within the multitasking environment of turn-taking, where halfway through an incoming turn a person is already planning a response. Figure 8.4 models this and is based on the processing of the "crunch zone," the latter part of an incoming turn (see Figure 8.1). Here, Bob is listening to Anne's turn; during the incoming turn, he first concentrates on comprehending what she is saying. As soon as he grasps the essential point or speech act, Bob calls on his production machinery and begins to formulate a response. This goes all the way down the chute to be clothed in the phonology and articulatory programs. Meanwhile, he is still listening to Anne and parsing the incoming turn, looking for points of possible syntactic completion, the end of a possible turn. As soon as one of these is detected, he checks for prosodic (or gestural) cues to turn closure. When such prosodic cues are detected, he launches his response. Given the natural limit to human response times, the response will emerge around 200 msec after the end of Anne's turn (as in the mode of the typical response times, shown in the inset histogram).

The upshot of this is that we may be deeply skeptical whether a machine can model all these processes in real time, in a way that could match a human interactant. "Deep learning" on vast databases of interactive discourse may help to capture interactive routines, but there is no prospect of a solution to the creative abduction of "the inference to the best explanation" under reflexive reasoning, the Gricean inference to intent, in the split-second manner typical of human interaction. Above all, we still have only the poorest descriptive

[4] This is the effect where visual lip reading interacts with the acoustic signal to produce a blended perception, allowing fruitful experimentation of multimodal processing. Another unsolved problem is what counts as synchronization with nested structures; for example, an indexical point with "Is that big red truck over there John's?"

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

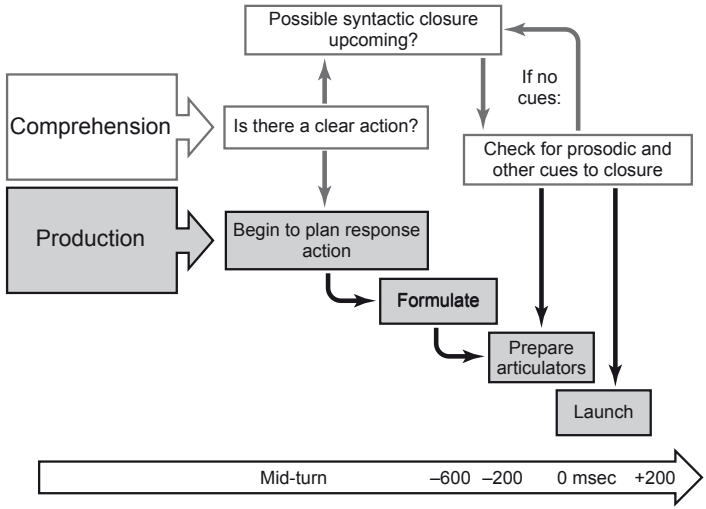


**Figure 8.4** How comprehension calls production midway through an incoming turn (from Levinson and Torreira 2015).

coverage, let alone understanding, of the many features that generate human communicative interaction.

*Fortunately, this does not spell doom for interactional robotics.* There are two saving graces. First, humans are good at adapting their interaction to interactants with lesser skills: to infants, children, foreign language learners, and even dogs. A very useful exercise would be to try to work out exactly which properties of human interaction are indispensable, and through which outputs these would best be embodied in an interacting machine (for suggestions, see Thomaz et al., this volume). Second, humans have a natural instinct to attribute deep interactive cognition to simple machines, lower animals, and even causal events without any cognition involved. It is this tendency that makes humans susceptible to superstition, religion, witchcraft, and animism (Levinson 1995, 2006). The very properties that make human interaction possible are freely read into the world around us. This accounts for some of the signal achievements in robotic interaction, as in the finding that autistic children are brought out of their interactive shells by playing with robots (for a review, see Diehl et al. 2012). It also accounts for our wonder, like Ovid's Pygmalion falling in love with his sculptural bride, at our abilities to make even shallow simulacra: the medieval fascination with automata (Truit 2015) or "Erica" in Ishiguro's Symbiotic Human–Robot Interaction Project (Glas et al. 2016). In restricted domains (e.g., travel agencies, directory enquiries), where specific human goals can be presumed, robotic performance can pass muster because we imagine greater competence than is actually present. Such bridgeheads will help to increase the utility of our machine companions, but we should not delude ourselves that we can really simulate the complex behavior that we

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

instinctively produce in our own actions, of which we have only the slightest analytical grasp.

## Conclusion

Human social interaction is our elite capacity. Other species have spectacular navigation abilities, the ability to sleep on the wing, or swim faster than a frigate, but our highly evolved trick is communicative interaction. This has made cultural transmission possible and has propelled us into the position of the dominant species on the planet. The ingredients of this elite capacity are scarcely understood, but they include capacities to model each other's plans and actions, to foresee them, and to plan ahead accordingly. Human communication is based on this: given the asymmetry in processing latencies in comprehension and production, the only way we can maintain the observed pace of communicative interaction is by predictive comprehension and preemptive production. The modeling of another's actions requires deep knowledge of the individual (in the case of known social others) and extensive cultural expectations, including interpersonal rituals in all cases. The speed and depth of computation is not likely to be matched by any machine in our lifetimes. Perhaps in some future world, it may be partially modeled by "breeding" machines using an analogue to unnatural selection, Darwin's selective breeding, to recapitulate the evolution of our elite capacity. In the meantime, human–machine interaction can trade on our charitable overattribution of interactional intelligence to anything that moves or squawks.

## Acknowledgments

I wish to thank Elena Lieven for early comments on a draft and to the members of my working group (Thomaz et al., this volume) for spirited discussion, to the reviewers who forced clarifications, and to Edith Sjoerdsma for her help with the manuscript. Finally, many thanks to Julia Lupp, Aimée Ducey-Gessner, Marina Turner, and Catherine Stephen for making participation a delight.

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