Great ape communication as contextual social inference: a computational modelling perspective

Manuel Bohn[1], Katja Liebal[2], Linda Oña[3], & Michael Henry Tessler[4,5]

[1] Department of Comparative Cultural Psychology, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

[2] Institute of Biology, Leipzig University, Leipzig, Germany

[3] Naturalistic Social Cognition Group, Max Planck Institute for Human Development, Berlin, Germany

[4] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, USA

[5] DeepMind, London, UK

Author Note

Correspondence concerning this article should be addressed to Manuel Bohn, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany. E-mail: manuel_bohn@eva.mpg.de

Abstract

Human communication has been described as a contextual social inference process. Research into great ape communication has been inspired by this view to look for the evolutionary roots of the social, cognitive, and interactional processes involved in human communication. This approach has been highly productive, yet it is often compromised by a too-narrow focus on how great apes use and understand individual signals. This paper introduces a computational model that formalises great ape communication as a multi-faceted social inference process that relies on information contained in the signal, the relationship between communicative partners, and the social context. This model makes accurate qualitative and quantitative predictions about real-world communicative interactions between semi-wild-living chimpanzees. When enriched with a pragmatic reasoning process, the model explains repeatedly reported differences between humans and great apes in the interpretation of ambiguous signals (e.g. pointing gestures). This approach has direct implications for observational and experimental studies of great ape communication and provides a new tool for theorising about the evolution of uniquely human communication.

Keywords: Communication, Primates, Social cognition, Evolution, Computational modeling

Great ape communication as contextual social inference: a computational modelling perspective

Introduction

When discussing the origins of human communication, Levinson and colleagues [1,2] introduced the idea of a human *interaction engine*. This metaphorical engine is assembled from a range of social-interactional parts that, when put together, enable uniquely human forms of communication, including conventional language. Each part was assumed to have deep roots in our evolutionary history and might therefore – in one form or the other – also be found in other primates. Inspired by these ideas, this paper introduces a computational model that specifies the role that social-interactional processes play in great ape and human communication.

What are the parts that the human interaction is built from? First and foremost, human communication is seen as intentional. Senders produce signals to convey intentions and receivers use these signals to infer the sender's intentions [3–6]. As such, communication is deeply linked to reasoning about mental states. Signals, including conventional language, are used to express intentions but the link between signals and intentions is not rigid. There is always residual ambiguity that requires communicators to make additional (pragmatic) inferences – a second key feature of human communication. Such inferences are licensed by a set of assumptions that humans hold about the nature of communication and social interaction more broadly. One such assumption is that communication occurs within some form of common ground – a shared body of knowledge and beliefs that builds up during social interaction and serves as the background against which signals are interpreted [7,8]. Another assumption is that communication is cooperative such that senders choose their signals so that the receiver is more likely to infer the underlying intention. The receiver takes this into account when interpreting the signal.

The engine assembled from these – and many other – parts is independent of any particular modality. Multi-modality is seen as the norm, not an exception in human communication. The system is also highly flexible. Sometimes a tiny hand gesture might be enough to get a message across; at other times, the same meaning might require a long, elaborate utterance comprised of multiple signals that are combined according to conventional rules (grammar). Or as Levinson and Holler [2] put it, "The system remains highly flexible, allowing us to shift the burden from words to gestures as required by the current communicative needs." Many roads lead to Rome in human communication and *what* works *when* depends on the social-interactional embedding. The system is also independent of the availability of conventional (or evolved) signals. Conventional language is assumed to rely on the engine in just the same way as non-conventional communication. New signals can be invented and understood on the spot and later even conventionalized into new languages [9–17].

The picture that emerges here provides an interesting starting point for an evolutionary research program because it decouples human communication from conventional language. The idea is that there is probably no direct link between the kind of signals our ancestors used (which might be comparable to what we see in great apes) and human language. The link lies in *how* signals are used, that is, the social and cognitive underpinnings of communication. Once the interaction engine was in place, our ancestors started using and creating signals that, via intermediate proto-languages, evolved to become what we today see as conventional language [18–22]. Thus, in addition to looking for structural features in animal communication that directly resemble aspects of conventional language (e.g. arbitrary sound-to-meaning mappings or combinatorial syntax [23–27]), comparative researchers can also ask which social-interactional processes underlie communication in other animals. In the next section, we will briefly summarize research in this tradition, with a focus on great ape communication.

A comparative approach to human language: The intentional nature of great ape communication

It is beyond the scope of this paper to give a comprehensive summary of existing research on primate communication. We will focus on two aspects that have received considerable attention in comparative research: signalers' intentional signal production and receivers' extraction of the intended meaning of a signal. We will show that research on these two aspects of great ape communication varies drastically depending on whether the focus is on vocal, gestural, or facial signals. To make matters worse, there are also marked differences between research on the production versus the perception or comprehension of signals.

To identify acts of intentional communication in great apes and other nonhuman primates, Leavens and colleagues [28] suggested a set of criteria derived from research on pre-linguistic communication in human infants [29]. These include the sender's sensitivity to the presence of other individuals, visual orienting behaviour and monitoring of the receiver, the adjustment of signal use to the receiver's attentional state, and the use of attention-getting behaviors if receivers are not visually attending. Finally, senders are expected to continue signaling and to elaborate signal use in case initial communicative attempts fail.

There is now ample evidence that great apes are intentional communicators in that sense, not only in the gestural modality [30,31]. For example, several species of great apes adjust their signal use to the attentional state of the receiver and only deploy visual gestures if the receiver is attending [28,32]. They also wait for a response and persist in their communicative attempts and might even elaborate their gesture use if the receiver does not react [28,33,34]. Sumatran orang-utans use gestures and also some facial expressions flexibly to achieve a variety of social goals [35,36]. Furthermore, wild chimpanzees are more likely to produce alarm calls when other individuals are unaware of a potential threat [37,38].

However, which and how many of the criteria for intentional communication are applied does not only vary across studies but also across modalities [30]. While intentional use is an integral part of defining a gesture, until more recently, this aspect was not considered important in vocal and facial research [39], resulting in the common but unjustified dichotomy between intentional gestures and emotional vocalizations and facial expressions [6].

The different theoretical and methodological approaches in vocal, gestural, and facial research have serious downstream consequences for research on primate communication more broadly. Gesture researchers focus on the behaviour of the sender because of the importance of intentional signal production, while vocal and to a lesser extent also facial researchers focus on signal perception and how receivers extract a signal's meaning. Vocal researchers, for example, frequently use playback experiments to study receivers' reactions to a very specific call to identify the meaning or function of this call [40]. As a consequence, vocal researchers are interested in context-specific signals, with very specific meanings, while gesture researchers investigate the flexible use of one signal across different contexts and argue that the information conveyed by a gesture might differ depending on the context in which it is used. Gesture researchers further largely ignore context-specific signals, as this would not fulfil the criterion of flexible usage, which is often considered an additional marker of intentional use [30,35].

Meaning is also conceptualized very differently across modalities, depending on whether the focus is on the signaler's or receiver's behaviour [39]. While gesture researchers focus on the message the signaler intends to communicate, vocal (and partly also facial researchers) focus on the 'meaning' extracted by the receiver [41,42]. As a consequence, it is difficult – if not impossible – to compare findings across modalities with regard to how nonhuman primates' communicative interactions are shaped by contextual information and how they ‘make sense’ of others’ communicative attempts.

Only more recently, there has been some cross-fertilization in both vocal and gesture research. Vocal researchers report that some vocalizations are less context-specific than previously thought [43], while gesture researchers started to assign specific meanings to individual gestures [44,45].

Despite these recent developments, it is important to highlight that research on primate communication has almost exclusively used a uni-modal approach: the majority of research focused either on gestural, vocal, or facial signals, and only very few studies investigated more than one signal modality simultaneously [46–49]. There are a number of different reasons why researchers artificially break up the communicative process into components and study each of them in isolation [50]. For example, researchers are trained in the theoretical approach and methods of their focal modality; methods used to study one modality (e.g., playback experiments) are not easily applicable to another modality.

There is, however, a deeper and more fundamental problem: we lack a theoretical account of how the different components integrate with one another. For human communication, Enfield [51], for example, proposed that composite utterances, incorporating multiple signals of multiple types, “[. . .] are interpreted through the recognition and bringing together of these multiple signs under a pragmatic unity heuristic or co-relevance principle, i.e. interpreter’s steadfast presumption of pragmatic unity despite semiotic complexity”. In other words, the recognition of each component’s (encoded) meaning is enriched by (the interpretation of) additional information, such as the meaning provided by the context in which this utterance is embedded. For primate communication, an equivalent theoretical account is yet missing and many of the following questions remain unsolved. How do different signals relate to one another? That is, how does the combination of a gesture with another signal (e.g. gesture, facial expression, or vocalization) change the meaning or usage of the initial gesture? What role does the social context play? Our goal for the rest of the paper is to sketch out such a theoretical account in the form of a computational model. As a first step, we will briefly introduce the Rational Speech Act (RSA) framework which formalises some of the reasoning processes implied by the interaction engine and from which we took inspiration.

Computational models of inferential communication in humans

A core challenge for a multi-layered, multi-modal system is to specify how the different information sources – the aspects of the utterance and the context that relate to the message being communicated – flow together [51–54]. The RSA framework sees communication as a socially guided inference process [55,56]. A hypothetical receiver in the model is assumed to reason about the intention that underlies the sender's production of an utterance in context[1]. Importantly, the receiver assumes that the sender is communicating in a cooperative way, choosing utterances that are maximally informative for the receiver given the context. This assumption allows the receiver to go beyond the literal meaning of the words that are used and to make pragmatic inferences.

The RSA framework has been successfully used to model a range of language understanding phenomena as pragmatic inferences including scalar and ad-hoc implicatures, non-literal language, politeness, and vagueness among others [55,57–61]. More recently, it has been used to predict how adults and children integrate different information sources to make inferences about what a sender is referring to [62]. In one study, Bohn and colleagues [63] measured children's developing sensitivity to different information sources, for example, their linguistic knowledge or their sensitivity to common ground. Then they used an RSA-type model to predict what should happen when people are confronted with multiple information sources at once. When they compared these

[1] The RSA framework usually uses *speaker* and *listener* to describe the agents involved. Here we continue to use the terms *sender* and *receiver* instead to be more inclusive of non-human and human multi-modal communication.

predictions to new experimental data, they saw a very close alignment between the two, both qualitative and quantitative. To learn more about the integration process itself, they formalized a range of alternative models that varied in their assumptions about which information sources children used and how they integrate them. They found that children's behaviour was best predicted by a model that assumed rational integration of all available information sources. Interestingly, the integration process was best described as stable across development. That is, even though children might change in how sensitive they are to different information sources, the way they integrate them seems to be stable across development. These studies illustrate how computational models can be used as a tool to study multi-layered communication.

For the model we describe below, we take inspiration from the RSA framework. The connection is mainly conceptual: we see communication as a socially guided inference process that relies on multiple, context-dependent information sources. There is, however, little structural overlap in terms of the implied cognitive mechanisms. In a later section, we explore how the social reasoning processes that are structural characteristics of RSA can be used to explain differences between great ape and human communication when it comes to interpreting novel and ambiguous signals.

Models of primate communication

Our main goal in this paper is to formulate a computational model of great ape communication. We focus on the in-the-moment comprehension of communicative acts. We ask how a receiver makes inferences about the intentions of a sender based on information contained in the signal, the relationship between communicative partners, and the social context. The process of in-the-moment comprehension has received little attention in previous modeling work in primate communication. We briefly review some of the earlier literature before laying out our approach.

Most formal work in primate communication has focused on modeling the production of different primate calls [64,65]. Though relevant for answering questions about the evolution of speech, this work does not help us understand the social-interactional nature of primate or ape communication. In a very ambitious project, Stuart Altmann[2] [66] used stochastic models to predict the socio-communicative behavior of rhesus monkeys (*Macaca mulatta*). He observed large groups of monkeys living on Cayo Santiago for two years with the goal to develop an ethogram of the species' social behavior. Next, he used his observations to define transitional probabilities between different behaviors. That is, he asked how well one can predict an individual's behavior if the previous behaviour (by the same or another individual) is known. He did this for pairs of behaviours, but also for longer sequences. Perhaps unsurprisingly, he found that the behavioural stream is not a random sequence of events, but that behaviours cluster in a systematic way. In a very broad sense, we take this as an inspiration to look for a wider set of determinants when trying to predict in-the-moment comprehension and reactions.

Arbib and colleagues [67–70] focused specifically on gestural communication. Their main goal, however, was to model the ontogeny of gestures. Their model shows how behavioural patterns can evolve into communicative gestures during direct, physical interaction. Given their specific aim, the authors saw the gesture as the sole cause of changes in the receiver's behaviour. Comprehension is treated as an associative learning process during which the observation of a particular action becomes paired with a particular reaction (i.e. change in the receiver's goal state). The result is a linear mapping between observing a gesture and producing an outcome. In our model, we loosen this assumption and take into account that multiple information sources influence the response to a gesture.

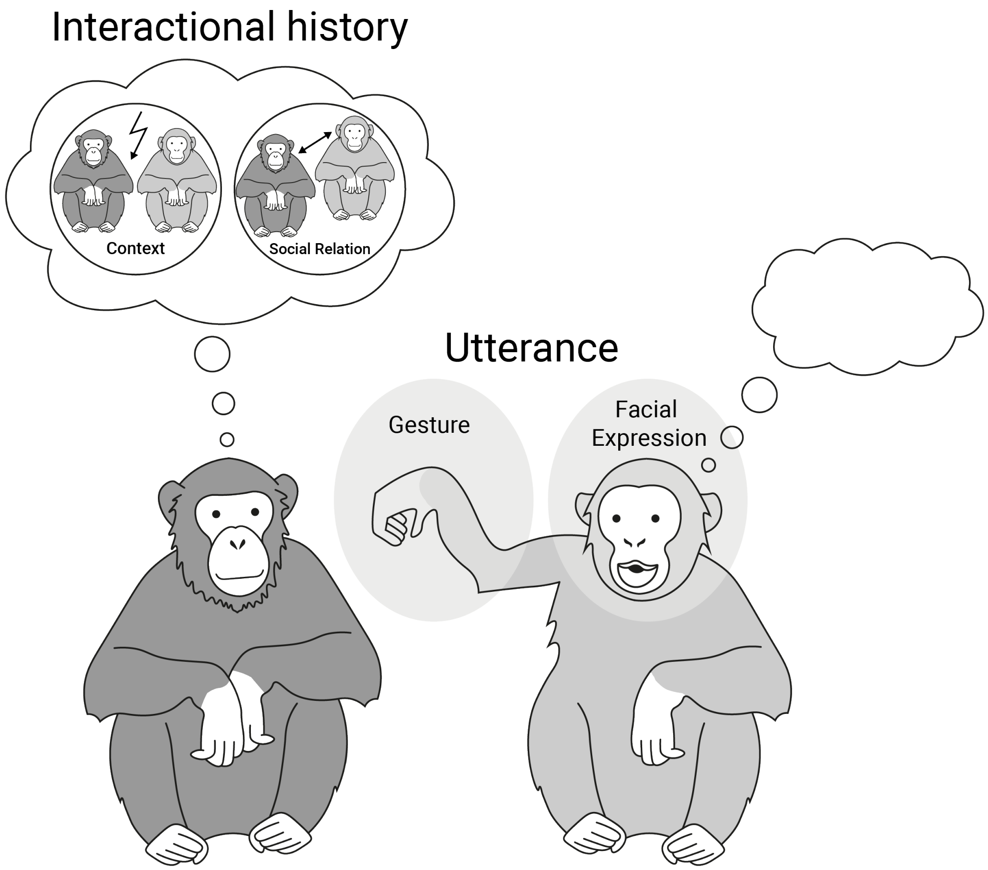

Figure 1. Schematic overview of the computational model. The sender (right) is producing an utterance and the receiver (left) tries to infer the intention of the sender based on the information sources available. The model takes in information provided by the utterance (gesture and facial expression) and the interactional history (immediate social context and dominance relation).

A computational model of chimpanzee communication

In this section, we introduce a Bayesian computational model of great ape communication. In contrast to standard statistical procedures (e.g. linear regression) which describe a particular data set, our model describes the inference processes we assume to underlie great apes' interpretation of communicative signals in context. These inference processes are built into the model structure and the model provides an account of the process that generated the data. Such a generative model can be used to predict and

[2] We are grateful to David Leavens for pointing us to Altmann's work.

explain data sets (see below), but its main purpose is to provide a theoretical account of the phenomenon in question. In what follows, we first present a very general formulation of our model and then further specify it to capture a particular type of communicative interaction. We then evaluate the model based on an existing data set.

We see great ape communication as a contextualized social inference problem. That is, the sender produces an utterance which the receiver uses to make inferences about the sender's intention (Figure 1). Utterances can be composed of different types of signals coming from different modalities (e.g. gestures, vocalizations, facial expressions, etc.). Inferences are contextualized in that, not just the utterance, but also the social context of the utterance as well as the relationship between the sender and receiver influence the receiver's interpretation. Thus, multiple information sources have to be integrated. We explore the hypothesis that this integration process occurs via a rational Bayesian procedure. This contrasts with the use of the term rational as describing a rule-based (i.e. logical) form of drawing conclusions. Here, we assume that the receiver's a posteriori belief is optimal given the receiver's prior beliefs and the constituent information sources they receive [71–73]. Given the simplicity of our model, we do not assume any limitations with respect to the cognitive resources our communicative agents have at their disposal. However, our approach could easily be extended in this direction, foe example, with resource-rational considerations [74]. The model is formally defined as

$$P(i \mid u) \propto P(u \mid i)P(i) \tag{1}$$

with $P(i \mid u)$ being the probability that the sender has intention i given utterance u. This decomposes into the likelihood of producing an utterance given an intention $P(u \mid i)$ (e.g. raising one's arm when wanting to be groomed) and the prior probability of having an intention in the first place $P(i)$ (e.g. wanting to be groomed). This very general formulation can be used as a framework to evaluate different hypotheses about which social information sources contribute to the likelihood and the prior; that is, which information sources play an important role in great ape communication.

Next, we spell out one variant of the model, which was in part determined by the data set that we had available for evaluation. As mentioned above, the general framework could be used with more, fewer, or different information sources. For the purpose of the current paper, the likelihood is defined by the semantics associated with a gesture, $\mathcal{L}(g, i)$, and a facial expression, $\mathcal{L}(f, i)$, which independently contribute to make up the utterance:

$$P(u \mid i) = P(g, f \mid i) = \mathcal{L}(g, i \mid \theta_g)\mathcal{L}(f, i \mid \theta_f) \tag{2}$$

Signals have "soft semantics", that is, in contrast to a truth-functional (Boolean) semantics, we assume a probabilistic mapping between a signal and an intention (defined by the parameters θ_g and θ_f) [75]; where θ_g is the strength of association between the gesture and the intention and θ_f that of the facial expression and the intention). The utterance is contextualized by the prior probability of the intention, $P(i)$, which we take to be a function of the context, and the social relation between individuals, $P(i \mid c, s)$:

$$P(i) = P(i \mid c, s) = \rho_c \rho_s \tag{3}$$

The direction and strength of the context and social relation components are defined by the parameters ρ_c and ρ_s (where ρ_c denotes the association between the context and the intention and ρ_s that between the social relation and the intention). In the example below, we provide more information about the interpretation of these parameters.

To evaluate the model, we used it to predict the outcome of communicative interactions between semi-wild-living chimpanzees (*Pan troglodytes*). The data is taken from the study by Oña and colleagues [49] in which the authors observed two groups of chimpanzees (72 individuals) living in the Chimfunshi Wildlife Orphanage Trust in Zambia.

They investigated if signal combinations were used in different contexts and/or elicited different responses compared to signals used alone. For every communicative interaction, they recorded the signals the sender produced, the context in which they were used and the reaction of the receiver. More specifically, they coded the type of manual gesture using a form-based coding scheme, differentiating between morphological configurations of the joints of the arm, hand, and fingers. Using this procedure, they identified two frequently occurring gesture types: *stretched-arm*, consisting of an extended arm with both the arm and hand stretched, and *bent-arm*, with either hand or forearm bent and the back of the hand or arm directed at the receiver. Facial expressions were coded using a modified version of the human Facial Action Coding Scheme (FACS)[76] developed to identify facial movements of chimpanzees (chimpFACS)[77].The *bared-teeth* face, with the mouth either closed or slightly opened, and the mouth corners laterally retracted and teeth fully exposed, was identified in addition to the *funneled-lip* face, consisting of an open, rounded mouth with protruded lips. When one of the gestures was combined with either of these facial expressions, this was considered a gesture-facial expression combination. When the gesture was used without a facial expression, the face was coded as neutral. Facial expressions produced in isolation, without an accompanying gesture, were not included. The social context of the interaction was coded as either positive (e.g. greeting, grooming, play), or negative (e.g., physical conflicts, harassment). The social relationship between the sender and receiver was considered by coding if signals were directed towards a lower or higher-ranking individual. Finally, the outcome of the interaction (i.e., the response of the receiver) was classified as either affiliative (receiver approaches the sender and shows behaviours such as embracing, grooming or play) or avoidant (receiver is avoiding or ignoring the sender, e.g., by turning away from, hitting or pushing the sender).

As noted above, in our model, the gesture and the facial expressions contribute to the utterance (the likelihood) and the social context and the relationship contribute to the prior. We assigned parameter values to each of the components of the communicative interactions. The goal was to show that by choosing intuitive parameter values, our model can give rise to the data we observed. These values range between 0 and 1 and represent the degree with which a component is indicative of a positive (affiliative; 0 - 0.5) or negative (avoidant; 0.5= 1) interpretation. We assumed the stretched-arm gesture to be weakly negative (θ_{gs}= 0.53) and the bent-arm gesture to be weakly positive (θ_{gb}= 0.47). Neutral facial expressions were set to be neutral (θ_{fn}= 0.5), bared-teeth expressions were set to be weekly negative ($\theta_{fb} = 0.6$), and funneled-lip expressions to be strongly negative ($\theta_{ff} = 0.9$). A negative context was set to be negative ($\rho_{cn} = 0.7$) and a positive to be positive ($\rho_{cp} = 0.3$). Finally, we assumed that a positive reaction was likely for a dominant sender ($\rho_{sd} = 0.25$) and a negative outcome likely for a subordinate sender ($\rho_{ss} = 0.75$).

We want to highlight that even though these parameter values are inspired by prior work and common sense, they are to some extent arbitrary and should not be taken to reflect a strong commitment to the role the individual components might play in a different context. Their main purpose is to capture the idea that different components of the communicative interaction are more or less associated with a particular response. Ideally – and hopefully in future work – these parameters would be directly estimated based on a training dataset and then used to predict a test dataset. Given the size of the dataset we had available, this approach was not possible here. The code that spells out the model architecture and the processing algorithms and that can be used to reproduce the results is available in the associated online repository: https://github.com/manuelbohn/RSApes.

Based on the model and the parameter settings we generated predictions for all possible combinations of gestures, facial expression, dominance relationship, and social context. We compared these predictions to the observations made by Oña and colleagues [49]. Our model makes predictions about the receiver's interpretation of the utterance in context. The data, however, only recorded the receivers' reactions – as interpreted by the human coders. We assume that the receiver's reaction is guided by their interpretation of the utterance: When inferring a negative intention, the receiver shows an avoidant reaction and when inferring a positive intention, they show an affiliative reaction. Thus, for the purpose of the model comparison, we assume a one-to-one mapping between the interpretation of the sender's message and the receiver's reaction.

Observations in the data were not equally distributed across all possible combinations. To evaluate the model predictions, we focused on combinations that had at least five observations. All combinations that fulfilled this criterion were observed in a negative social context. When we compare the model predictions to the data, we therefore only visualize the negative context (Figure 2). Note, however, that our model also generated predictions for the positive context.

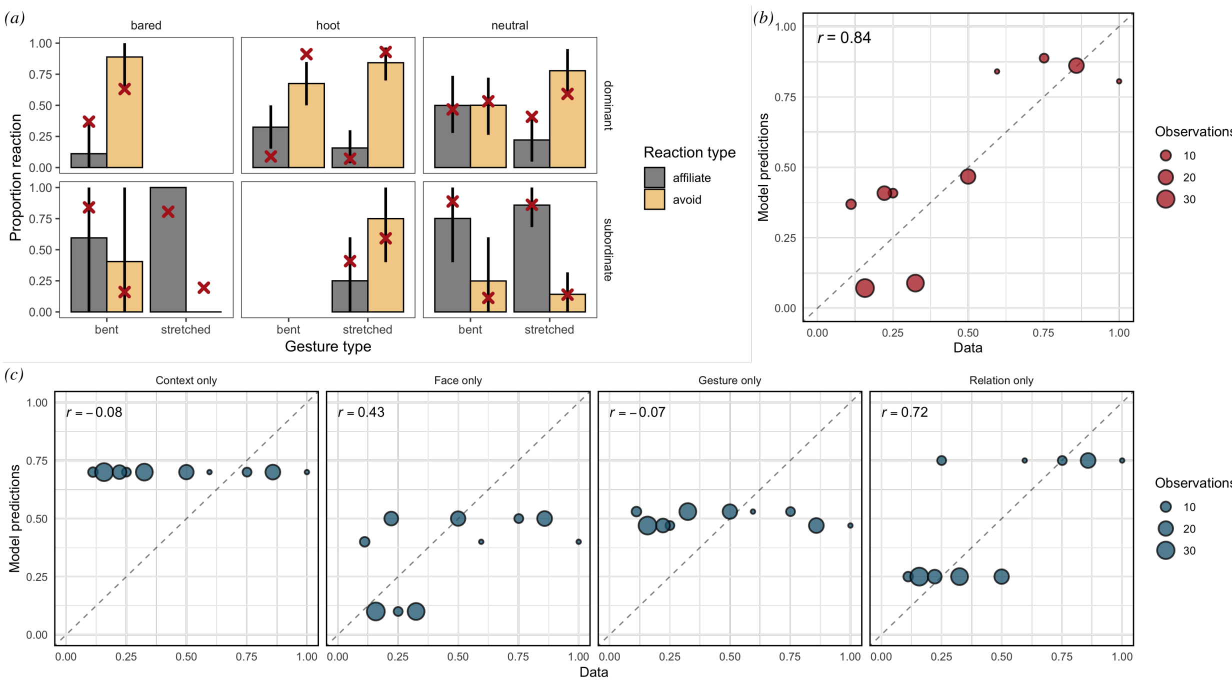

Figure 2. Model predictions compared to data from [49]. Panel (a) shows the mean proportion (bars) of affiliative and avoidant reactions for combinations of gesture, facial expression, relationship, and social context in the data. Only combinations with more than 5 observations are shown. Error bars are 95% Confidence Intervals based on a non-parametric bootstrap. Red crosses show model predictions. Panel (b) shows correlations between model prediction and data for avoidant reactions. The size of each point is proportional to the number of observations for a particular combination in the data. Panel (c) shows correlations for reduced models that focus only on a single component (with all other parameters set to 0.5).

In Figure 2, we can see that the full model explains the data well, both quantitatively and qualitatively. The model predictions go in the same qualitative direction as the data, predicting more negative reactions when more were observed. Furthermore, many of the model predictions also align quantitatively with the data, resulting in a high correlation between the two (Figure 2b). Let us take a closer look at some of these patterns. In most cases, the qualitative pattern in the data was the same for both gesture types. For example, in a negative context (Figure 2 only includes the negative context), with a subordinate sender and a neutral facial expression, no matter if a bent or a stretched-arm gesture was used, there were more affiliative reactions. Our model predicts this pattern despite the fact that we took the stretched-arm gesture to be associated with a negative intention. The reason for this is that both gestures were assumed to have weak meanings. As a consequence, they had very little predictive power when a different, stronger information source (the dominance relationship in this case) was also available.

Next, we used this modeling framework to illustrate the theoretical point made above, namely that a focus on a single aspect of great ape communication is likely to yield an incomplete picture of the interaction. We formulated four reduced models, which use the same parameter settings as above, but selectively focused only on one of the components (all other parameters set to 0.5). When comparing the predictions from these reduced models to the data, we saw that none of them captured the data equally well compared to the full model (Figure 2c)[3]. For example, the models focusing only on the context or the gesture completely fail to capture any structure in the data. These results, however, should be taken with a grain of salt given the – rather arbitrary – way in which we chose the parameter values. Nevertheless, we think the the results nicely illustrate how computational modeling can be used as a powerful tool to study great ape communication. In the next section, we explore ways in which we can use this tool to theorize about some

[3] In the online repository, we also include a model in which the strength of the meaning of gestures and facial expressions was switched. That is, gestures were assumed to have a rather strong meaning and facial expressions a weak one. This model makes worse qualitative and quantitative predictions compared to one presented in the paper.

potential differences between ape and human communication.

Pragmatics as an amplifier

In their description of the interaction engine, Levinson and Holler [2] point out that "language is the tip of an iceberg riding on a deep infrastructure of communicational abilities". Part of this deep infrastructure is pragmatics. As noted in the introduction, the central idea is that utterances are not interpreted at face value, but that receivers go beyond the literal and make inferences about why the sender produced a particular utterance in context. A cornerstone of this reasoning is the assumption that the sender is cooperative and informative; they produce utterances that help the receiver to infer their intention.

In the following, we enrich our model of great ape communication by pragmatics – i.e., cooperative social reasoning. From an evolutionary perspective, we may say that our great ape model stands in for the last common ancestor of great apes and humans. To recapitulate, we assume that this ancestor (and modern great apes) rationally integrated different information sources to make inferences about the sender's intentions. This includes information contained in the utterance as well as the social context and the relationship between communicators. The pragmatic abilities are built on top of this basic infrastructure to provide modern human communication.

To evaluate this pragmatically enriched model, we want to focus on some peculiar differences that have been reported for the communicative abilities of great apes and humans. Numerous studies have shown that great apes struggle to spontaneously understand ambiguous signals, for example, pointing or novel iconic gestures [9,78–86] (with some particular exceptions [87,88]). That is, when confronted with a novel gesture or a new context, great apes usually fail to spontaneously use the gesture. These findings are peculiar because these gestures are naturally meaningful in that they either index (pointing) or resemble (iconic gestures) the referent. What is more, human children understand them spontaneously already very early in life [89–91]. Apes also seem to be somewhat sensitive to the natural meaning of these gestures. In the case of pointing, they often look in the direction the experimenter is pointing [92]. And in one study, iconic gestures were learned faster compared to arbitrary ones [93].

Why do apes struggle with spontaneous comprehension of these gestures? The results of the model above can be taken to suggest that the social context and the relationship between sender and receiver play an important role in great ape communication. In the experimental setups of studies on pointing or iconic gesture comprehension, these components are controlled for and therefore offer no information about the sender's intention [9,81,84]. Great apes are left with only the gesture. If that gesture was initially only vaguely associated with one or the other outcome, it would not provide sufficient information for apes to infer the sender's intention and thus to systematically select the referred-to object.

Why do humans spontaneously understand these gestures? We think that the notion of pragmatics as spelled out above can act as an amplifier of vague literal meanings. That is, a human receiver assumes that the sender produced a particular gesture in a cooperative and informative manner to inform them about their intention. The additional social reasoning singles out the gesture as a communicative act that was produced with the sole purpose to express a given intention (Figure 3. This line of argument is of course reminiscent of the idea that humans – but not great apes – are sensitive to cooperative communicative intentions [6]. However, we assume that pragmatic inferences just one information source that can be exploited and that they are graded – not all or nothing. Taken together, the degree to which pragmatic reasoning amplifies a meaning depends on a) the presence of a social reasoning mechanisms and b) expectations about how cooperative the sender is. Next, we substantiate these ideas via our modeling framework.

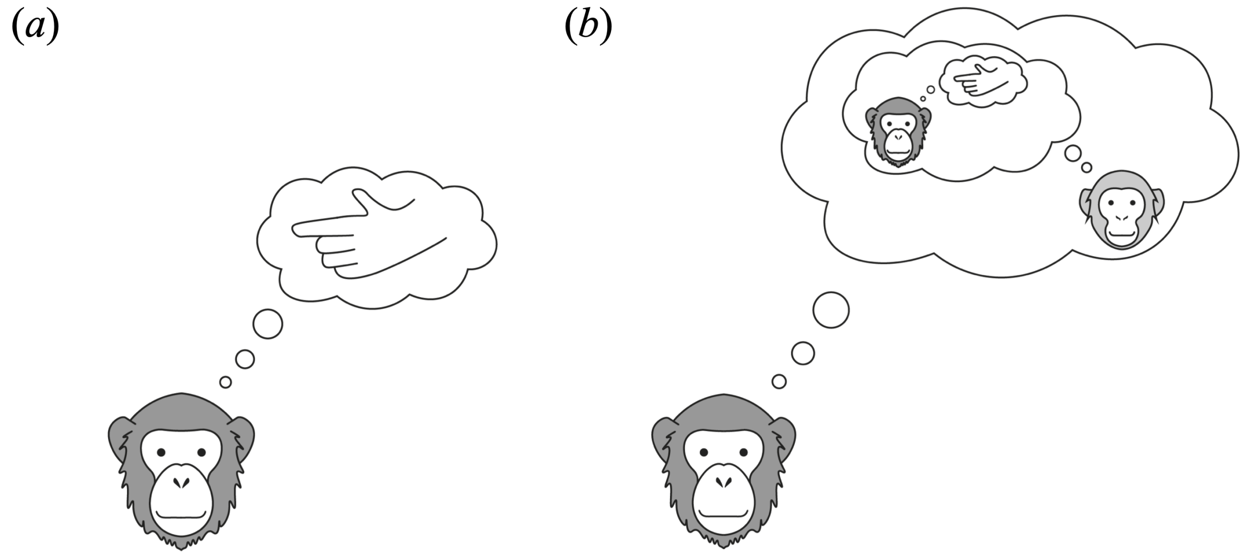

Figure 3. Schematic depiction of the added pragmatic reasoning component. The literal receiver (a) only reasons about the gesture whereas the pragmatic receiver (b) reasons about why the sender produced that particular gesture. The pragmatic receiver further expects the sender to produce the gesture with the goal of being informative.

The RSA framework introduced above is built around the assumptions that a) receiver reason about why senders produce certain utterances and b) receivers assume that senders communicate in a cooperative and informative way. This social reasoning component is formalized by embedding the model of the (zero-order) literal receiver (short-hand notation: P_{R_0}), in a model of the sender,P_{S_1}. This *pragmatic* sender chooses utterances so that they are informative for the literal receiver, while the literal receiver simply interprets utterances in line with their literal semantics. This literal receiver behaves exactly like in the great ape model (Figure 3). This illustrates the way in which our model of human communication is built around our model of great ape communication. At the highest level, we now have a pragmatic receiver, P_{R_1}. These additions change our model like so:

$$P_{R_1}(i \mid u) \propto P_{S_1}(u \mid i)P(i) \tag{4}$$

$$P_{S_1}(u \mid i) \propto P_{R_0}(i \mid u)^{\alpha_i} \tag{5}$$

$$P_{R_0}(i \mid u) \propto \mathcal{L}(u, i \mid \theta_u) \tag{6}$$

Equation (5) above shows that the degree of how informative the sender is assumed to be depends on the parameter α. The higher α, the more informative the sender is assumed to be. The effect of α, however, depends on the presence of the sender model, which represents the additional social reasoning component that we think is characteristic of human communication.

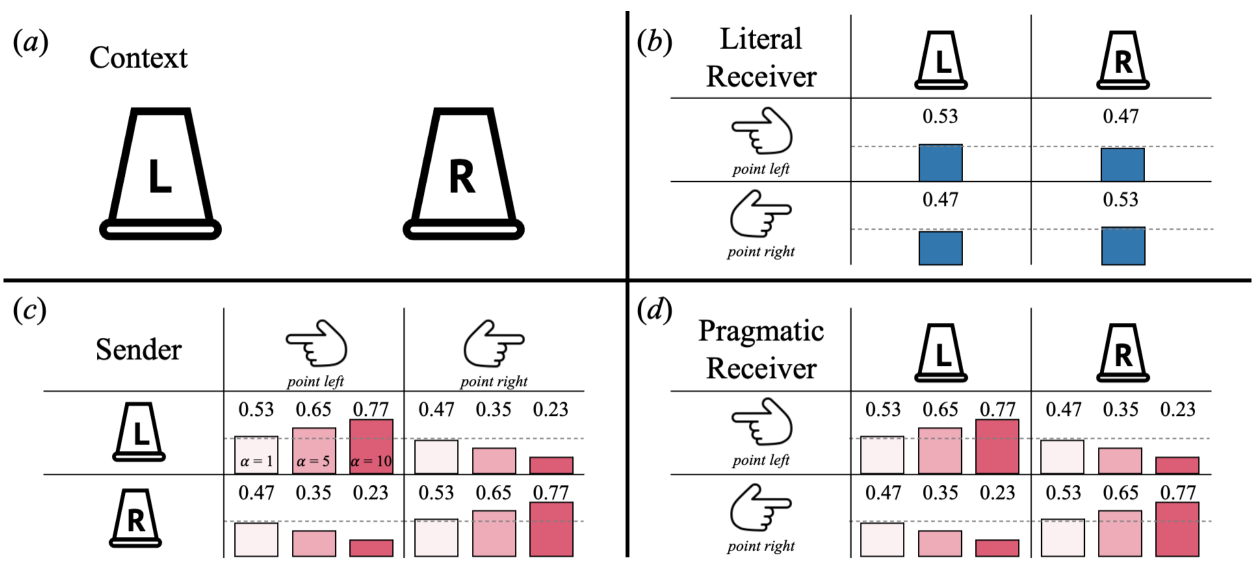

Figure 4. Application of the pragmatically enriched model to an object-choice task with pointing gestures. Panel (a) shows the context with the two locations (L = left and R = right) that can be referred to. Panel (b) gives the interpretation probabilities of a literal receiver. Panel (c) shows the production probabilities for the pragmatic sender for values of α = 1, 5, and 10. Panel (d) shows the interpretation probabilities of the pragmatic sender based on the production probabilities in panel (c). Colored bars visualize the probabilities in reference to chance (grey dashed line). Different shades of red in (c) and (d) correspond to the magnitude of α.

When we adopt such a model to a situation in which the receiver is faced with a vaguely meaningful gesture (e.g. a point or an iconic gesture; $\theta_u = 0.53$) without any additional contextual information, we see that the literal interpretation of the gesture simply reflects this vague meaning (Figure 4b). We also see that pragmatic reasoning amplifies the initially vague meaning (Figure 4d). As noted above, this is not due to the additional social reasoning component alone but critically depends on the receiver's expectation about cooperative communication (the parameter α, Figure 4c). This highlights the graded relation between assumptions about cooperativeness and pragmatic inference. Once again, we would like to point out that the specific parameter values we picked here are arbitrary and do not reflect a strong commitment to how great apes or humans interpret pointing gestures. They simply serve to illustrate the point that pragmatics may amplify vague natural meanings.

Implications and future directions

With the modeling exercise presented above we had two overarching goals. The first was to show that great ape communication is best thought of (and studied) as a multi-faceted, multi-modal, social inference process. We saw that the outcome of a communicative interaction was best predicted when signals, as well as contextual components, were taken into account. We do not say that studying these components in isolation is fruitless, but we do emphasize that focusing exclusively on, for example, the gesture or vocalization produced makes it less likely to understand the interaction that is unfolding. From our perspective, the different components play complementary roles in an integrated inference process.

Our hope is that our model proves to be a useful tool – or at least an inspiration – for future research. The approach by Oña and colleagues [49], in which many different aspects of a communicative interaction are coded, seems to be especially promising. Such work could easily be done using already existing video recordings. Models like the one presented here could then be used to specify how the different components work together. In addition, our framework provides a new way to test competing hypotheses. Instead of relying on qualitative predictions, alternative hypotheses can be formalized as alternative models and then directly compared in a quantitative way. Across studies, it would be interesting to see if general patterns emerge. For example, models that emphasize social-contextual components could make better predictions compared to models emphasizing information provided by the utterance. Or models prioritizing facial expressions could be found to outcompete models that more strongly emphasize gestures. Or vice versa in both cases. Experimental studies could gradually vary the information provided by signals and the social context to study how they trade-off with one another. Such an approach might reveal quantitative differences between humans and other primates where we currently assume qualitative ones. In all of this, we think that the study of great ape communication would benefit from an interdisciplinary approach in which computational modelers work together with primatologists and comparative psychologists. Hopefully, this will allow the field to move away from asking somewhat artificial questions about the importance of individual gestures, facial expressions or vocalizations and instead move towards more comprehensive theories of the actual processes that underlie communicative interactions.

We see our model as a first step that needs to be expanded in the future. The process that we capture in our model is in-the-moment comprehension, which is only a part of communicative interaction. An easy extension would be to look at the sender: We assume our model to be symmetric and so it could be easily used to generate predictions about what types of gestures, facial expressions and vocalizations the sender should produce in different contexts given the intention they want to communicate. Furthermore, it would be interesting to extend our model to capture the temporal dynamics of communication – that is, to include mechanisms that are used to clarify or emphasize a message. Candidate behaviours in primates could be acts of persistence, repetition or elaboration which are often seen in naturalistic and experimental settings [28,34]. Including this aspect might have consequences for the cognitive architecture of the model. For example, Arkel and colleagues [94] have suggested that a simple repair mechanism drastically changes the computational demands in human communication.

Our second goals was to demonstrate how pragmatic reasoning can act as a gradual amplifier for signals with vague meanings. This perspective might be helpful for theorizing about the gradual transition from animal to human communication. For example, Sterelny [21] has argued that the transition from animal to human communication involved shifting from code-based to ostensive inferential communication [21,95]. During this process, the tight signal-response coupling characteristic for code-based communication was loosened. This brought an increase in flexibility, allowing senders to use the same signal for different and potentially novel purposes. However, it also introduced ambiguity to the signal, which, according to Sterelny, was compensated by relying on social reasoning processes. This transition shifted the locus of selection from specific signal-response couplings to communicative behavior more broadly, with downstream consequences for other forms of cooperative interaction [96]. Our model formalizes the trade-off between ambiguity in the signal – which is characteristic of human communication [20,97]– and social reasoning. As such, it could be used as a starting point to formalize the gradual evolution of human ostensive-inferential communication.

The gradual emergence of pragmatic social reasoning in the evolution of human communication might have had further downstream consequences for the emergence of conventional communication systems. Recently, Hawkins and colleagues [98] embedded an RSA model of pragmatic in-the-moment inferences in a model of convention formation and showed how signals with vague meanings can give rise to conventional communication systems. The meaning of a signal can get fixed (e.g., further amplified) when it is repeatedly used within dyadic communicative interactions. Conventions form when partner-specific communicative conventions are gradually transferred, via a hierarchical Bayesian model, to novel communicative partners. Work by Woensdregt and colleagues [99] suggests that the presence of conventional communication systems further facilitates in-the-moment inferences about communicative intentions, leading to a cascading co-evolution of conventional communication systems and social reasoning.

Finally, our modeling approach informs discussions about the modality in which human language has evolved. For decades, there has been a strong divide between researchers arguing for a vocal or a gestural origin of language [19,46,50,100]. Recently, the idea that language origins were multi-modal has gained traction [46,100]. Our model provides a way of thinking about multi-modal communication. The model does not make any principled distinction between different modalities: for every signal, it simply asks how indicative it is for different intentions the sender might have. This explains how different signals influence each other during in-the-moment comprehension and could also be used to investigate how the burden may have shifted between modalities during the course of evolution.

Conclusion

Inspired by work on the human interaction engine, we have described a computational approach for how to study great ape communication in context. Our model assumes that great apes rationally integrate different information sources to make inferences about the intention behind a sender's utterance in context. Using existing data, we have shown that our model makes accurate predictions about the outcome of multi-modal communicative interactions between chimpanzees in different social contexts. Based on the idea that pragmatic reasoning – social reasoning paired with assumptions about cooperative communication – acts as an amplifier for vague meanings, we suggested an explanation for some peculiar differences between the way that great apes and humans interpret ambiguous signals. This approach illustrates some deep similarities between human and great apes communication, but also specifies in what way the human interaction engine might be equipped with some special parts.

References

1.

Levinson SC. 2006 On the human ‘interactional engine’. In *Roots of human sociality: culture, cognition and interaction* (eds N Enfield, S Levinson), pp. 39–69. Oxford : Berg.

2.

Levinson SC, Holler J. 2014 The origin of human multi-modal communication. *Philosophical Transactions of the Royal Society B: Biological Sciences* **369**, 20130302.

3.

Grice HP. 1991 *Studies in the way of words.* Cambridge, MA: Harvard University Press.

4.

Levinson SC. 2000 *Presumptive meanings: The theory of generalized conversational implicature.* Cambridge, MA: MIT press.

5.

Sperber D, Wilson D. 2001 *Relevance: Communication and cognition.* 2nd edn. Cambridge, MA: Blackwell Publishers.

6.

Tomasello M. 2008 *Origins of human communication.* Cambridge, MA: MIT press.

7.

Bohn M, Köymen B. 2018 Common ground and development. *Child Development Perspectives* **12**, 104–108.

8.

Clark HH. 1996 *Using language.* Cambridge: Cambridge University Press.

9.

Bohn M, Call J, Tomasello M. 2019 Natural reference: A phylo-and ontogenetic perspective on the comprehension of iconic gestures and vocalizations. *Developmental Science* **22**, e12757.

10.

Bohn M, Kachel G, Tomasello M. 2019 Young children spontaneously recreate core properties of language in a new modality. *Proceedings of the National Academy of Sciences* **116**, 26072–26077.

11.

Brentari D, Goldin-Meadow S. 2017 Language emergence. *Annual Review of Linguistics* **3**, 363–388.

12.

Fay N, Lister CJ, Ellison TM, Goldin-Meadow S. 2014 Creating a communication system from scratch: Gesture beats vocalization hands down. *Frontiers in Psychology* **5**, 354.

13.

Goldin-Meadow S, Feldman H. 1977 The development of language-like communication without a language model. *Science* **197**, 401–403.

14.

Sandler W, Meir I, Padden C, Aronoff M. 2005 The emergence of grammar: Systematic structure in a new language. *Proceedings of the National Academy of Sciences* **102**, 2661–2665.

15.

Senghas A, Kita S, Özyürek A. 2004 Children creating core properties of language: Evidence from an emerging sign language in nicaragua. *Science* **305**, 1779–1782.

16.

Lister CJ, Burtenshaw T, Walker B, Ohan JL, Fay N. 2021 A cross-sectional test of sign creation by children in the gesture and vocal modalities. *Child Development* **92**, 2395–2412.

17.

Kempe V, Gauvrit N, Panayotov N, Cunningham S, Tamariz M. 2021 Amount of learning and signal stability modulate emergence of structure and iconicity in novel signaling systems. *Cognitive Science* **45**, e13057.

18.

Bar-on D. 2013 Origins of meaning: Must we 'go gricean'? *Mind & Language* **28**, 342–375.

19.

Fitch WT. 2010 *The evolution of language*. Cambridge University Press.

20.

Pleyer M. 2017 Protolanguage and mechanisms of meaning construal in interaction. *Language Sciences* **63**, 69–90.

21.

Sterelny K. 2017 From code to speaker meaning. *Biology & Philosophy* **32**, 819–838.

22.

Planer R, Sterelny K. 2021 *From signal to symbol: The evolution of language*. MIT Press.

23.

Arbib MA, Liebal K, Pika S. 2008 Primate vocalization, gesture, and the evolution of human language. *Current Anthropology* **49**, 1053–1076.

24.

Arnold K, Zuberbühler K. 2006 Semantic combinations in primate calls. *Nature* **441**, 303–303.

25.

Boë L-J, Fagot J, Perrier P, Schwartz J-L. 2017 *Origins of human language: Continuities and discontinuities with nonhuman primates.* Peter Lang International Academic Publishers.

26.

Fedurek P, Slocombe KE. 2011 Primate vocal communication: A useful tool for understanding human speech and language evolution? *Human Biology* **83**, 153–173.

27.

Zuberbühler K. 2005 The phylogenetic roots of language: Evidence from primate communication and cognition. *Current Directions in Psychological Science* **14**, 126–130.

28.

Leavens DA, Russell JL, Hopkins WD. 2005 Intentionality as measured in the persistence and elaboration of communication by chimpanzees (pan troglodytes). *Child Development* **76**, 291–306.

29.

Bates E, Benigni L, Bretherton I, Camaioni L, Volterra V. 1979 *The emergence of symbols: Cognition and communication in infancy.* New York: Academic Press.

30.

Liebal K, Waller BM, Burrows AM, Slocombe KE. 2014 *Primate communication: A multimodal approach.* Cambridge University Press.

31.

Townsend SW *et al.* 2017 Exorcising grice's ghost: An empirical approach to studying intentional communication in animals. *Biological Reviews* **92**, 1427–1433.

32.

Poss SR, Kuhar C, Stoinski TS, Hopkins WD. 2006 Differential use of attentional and visual communicative signaling by orangutans (pongo pygmaeus) and gorillas (gorilla gorilla) in response to the attentional status of a human. *American Journal of Primatology* **68**, 978–992.

33.

Cartmill EA, Byrne RW. 2007 Orangutans modify their gestural signaling according to their audience's comprehension. *Current Biology* **17**, 1345–1348.

34.

Roberts AI, Vick S-J, Buchanan-Smith HM. 2013 Communicative intentions in wild chimpanzees: Persistence and elaboration in gestural signalling. *Animal Cognition* **16**, 187–196.

35.

Call J, Tomasello M. 2007 *The gestural communication of apes and monkeys*. New York: Lawrence Erlbaum Associates.

36.

Liebal K, Pika S, Tomasello M. 2006 Gestural communication of orangutans (pongo pygmaeus). *Gesture* **6**, 1–38.

37.

Crockford C, Wittig R, Mundry R, Zuberbühler K. 2012 Wild chimpanzees inform ignorant group members of danger. *Current Biology* **22**, 142–146.

38.

Crockford C, Wittig R, Zuberbühler K. 2017 Vocalizing in chimpanzees is influenced by social-cognitive processes. *Science Advances* **3**.

39.

Liebal K, Oña L. 2018 Mind the gap–moving beyond the dichotomy between intentional gestures and emotional facial and vocal signals of nonhuman primates. *Interaction Studies* **19**, 121–135.

40.

Fischer J, Noser R, Hammerschmidt K. 2013 Bioacoustic field research: A primer to acoustic analyses and playback experiments with primates. *American Journal of Primatology* **75**, 643–663.

41.

Font E, Carazo P. 2010 Animals in translation: Why there is meaning (but probably no message) in animal communication. *Animal Behaviour* **80**, e1.

42.

Smith WJ. 1965 Message, meaning, and context in ethology. *The American Naturalist* **99**, 405–409.

43.

Wheeler BC, Fischer J. 2012 Functionally referential signals: A promising paradigm whose time has passed. *Evolutionary Anthropology: Issues, News, and Reviews* **21**, 195–205.

44.

Hobaiter C, Byrne RW. 2014 The meanings of chimpanzee gestures. *Current Biology* **24**, 1596–1600.

45.

Graham KE, Hobaiter C, Ounsley J, Furuichi T, Byrne RW. 2018 Bonobo and chimpanzee gestures overlap extensively in meaning. *PLOS Biology* **16**, e2004825.

46.

Slocombe KE, Waller BM, Liebal K. 2011 The language void: The need for multimodality in primate communication research. *Animal Behaviour* **81**, 919–924.

47.

Wilke C, Kavanagh E, Donnellan E, Waller BM, Machanda ZP, Slocombe KE. 2017 Production of and responses to unimodal and multimodal signals in wild chimpanzees, pan troglodytes schweinfurthii. *Animal Behaviour* **123**, 305–316.

48.

Hobaiter C, Byrne RW, Zuberbühler K. 2017 Wild chimpanzees' use of single and combined vocal and gestural signals. *Behavioral Ecology and Sociobiology* **71**, 1–13.

49.

Oña LS, Sandler W, Liebal K. 2019 A stepping stone to compositionality in chimpanzee communication. *PeerJ* **7**, e7623.

50.

Liebal K, Slocombe KE, Waller BM. 2022 The language void 10 years on: Multimodal primate communication research is still uncommon. *Ethology Ecology & Evolution*, 1–14.

51.

Enfield NJ. 2009 *The anatomy of meaning: Speech, gesture, and composite utterances.* Cambridge University Press.

52.

Holler J, Levinson SC. 2019 Multimodal language processing in human communication. *Trends in Cognitive Sciences* **23**, 639–652.

53.

Vigliocco G, Perniss P, Vinson D. 2014 Language as a multimodal phenomenon: Implications for language learning, processing and evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences* **369**, 20130292.

54.

Cavicchio F, Dachkovsky S, Leemor L, Shamay-Tsoory S, Sandler W. 2018 Compositionality in the language of emotion. *PloS ONE* **13**, e0201970.

55.

Frank MC, Goodman ND. 2012 Predicting pragmatic reasoning in language games. *Science* **336**, 998–998.

56.

Goodman ND, Frank MC. 2016 Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences* **20**, 818–829.

57.

Goodman ND, Stuhlmüller A. 2013 Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science* **5**, 173–184.

58.

Kao JT, Wu JY, Bergen L, Goodman ND. 2014 Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences* **111**, 12002–12007.

59.

Lassiter D, Goodman ND. 2017 Adjectival vagueness in a bayesian model of interpretation. *Synthese* **194**, 3801–3836.

60.

Tessler MH, Goodman ND. 2019 The language of generalization. *Psychological Review* **126**, 395.

61.

Yoon EJ, Tessler MH, Goodman ND, Frank MC. 2020 Polite speech emerges from competing social goals. *Open Mind* **4**, 71–87.

62.

Bohn M, Tessler MH, Merrick M, Frank MC. 2022 Predicting pragmatic cue integration in adults' and children's inferences about novel word meanings. *Journal of Experimental Psychology: General*

63.

Bohn M, Tessler MH, Merrick M, Frank MC. 2021 How young children integrate information sources to infer the meaning of words. *Nature Human Behaviour* **5**, 1046–1054.

64.

De Boer B, Tecumseh Fitch W. 2010 Computer models of vocal tract evolution: An overview and critique. *Adaptive Behavior* **18**, 36–47.

65.

Riede T, Bronson E, Hatzikirou H, Zuberbühler K. 2005 Vocal production mechanisms in a non-human primate: Morphological data and a model. *Journal of Human Evolution* **48**, 85–96.

66.

Altmann SA. 1965 Sociobiology of rhesus monkeys. II: Stochastics of social communication. *Journal of Theoretical Biology* **8**, 490–522.

67.

Arbib M, Ganesh V, Gasser B. 2014 Dyadic brain modelling, mirror systems and the ontogenetic ritualization of ape gesture. *Philosophical Transactions of the Royal Society B: Biological Sciences* **369**, 20130414.

68.

Gasser B, Arbib M. 2019 A dyadic brain model of ape gestural learning, production and representation. *Animal Cognition* **22**, 519–534.

69.

Arbib MA. 2016 Towards a computational comparative neuroprimatology: Framing the language-ready brain. *Physics of Life Reviews* **16**, 1–54.

70.

Gasser B, Cartmill EA, Arbib MA. 2014 Ontogenetic ritualization of primate gesture as a case study in dyadic brain modeling. *Neuroinformatics* **12**, 93–109.

71.

Yuille A, Kersten D. 2006 Vision as bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences* **10**, 301–308.

72.

Griffiths TL, Tenenbaum JB. 2006 Optimal predictions in everyday cognition. *Psychological science* **17**, 767–773.

73.

Chater N, Oaksford M. 1999 Ten years of the rational analysis of cognition. *Trends in Cognitive Sciences* **3**, 57–65.

74.

Lieder F, Griffiths TL. 2020 Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences* **43**.

75.

Degen J, Hawkins RD, Graf C, Kreiss E, Goodman ND. 2020 When redundancy is useful: A bayesian approach to 'overinformative' referring expressions. *Psychological Review* **127**, 591.

76.

Ekman P, Friesen WV. 1978 Facial action coding system. *Environmental Psychology & Nonverbal Behavior*

77.

Parr LA, Waller BM, Vick SJ, Bard KA. 2007 Classifying chimpanzee facial expressions using muscle action. *Emotion* **7**, 172.

78.

Bohn M, Kordt C, Braun M, Call J, Tomasello M. 2020 Learning novel skills from iconic gestures: A developmental and evolutionary perspective. *Psychological Science* **31**, 873–880.

79.

Dezecache G, Bourgeois A, Bazin C, Schlenker P, Chemla E, Maille A. 2019 Orangutans' comprehension of zoo keepers' communicative signals. *Animals* **9**, 300.

80.

Hare B, Tomasello M. 2004 Chimpanzees are more skilful in competitive than in cooperative cognitive tasks. *Animal behaviour* **68**, 571–581.

81.

Herrmann E, Tomasello M. 2006 Apes' and children's understanding of cooperative and competitive motives in a communicative situation. *Developmental Science* **9**, 518–529.

82.

Kirchhofer KC, Zimmermann F, Kaminski J, Tomasello M. 2012 Dogs (canis familiaris), but not chimpanzees (pan troglodytes), understand imperative pointing. *PLOS One* **7**, e30913.

83.

Tomasello M, Call J, Gluckman A. 1997 Comprehension of novel communicative signs by apes and human children. *Child Development*, 1067–1080.

84.

Tempelmann S, Kaminski J, Liebal K. 2013 When apes point the finger: Three great ape species fail to use a conspecific's imperative pointing gesture. *Interaction Studies* **14**, 7–23.

85.

Moore R, Call J, Tomasello M. 2015 Production and comprehension of gestures between orang-utans (pongo pygmaeus) in a referential communication game. *PLOS One* **10**, e0129726.

86.

Margiotoudi K, Bohn M, Schwob N, Taglialatela J, Pulvermüller F, Epping A, Schweller K, Allritz M. 2022 Bo-NO-bouba-kiki: Picture-word mapping but no spontaneous sound symbolic speech-shape mapping in a language trained bonobo. *Proceedings of the Royal Society B* **289**, 20211717.

87.

Mulcahy NJ, Call J. 2009 The performance of bonobos (pan paniscus), chimpanzees (pan troglodytes), and orangutans (pongo pygmaeus) in two versions of an object-choice task. *Journal of Comparative Psychology* **123**, 304.

88.

Lyn H, Russell JL, Hopkins WD. 2010 The impact of environment on the comprehension of declarative communication in apes. *Psychological Science* **21**, 360–365.

89.

Behne T, Carpenter M, Tomasello M. 2005 One-year-olds comprehend the communicative intentions behind gestures in a hiding game. *Developmental Science* **8**, 492–499.

90.

Rüther J, Liszkowski U. 2020 Ontogenetic emergence of cognitive reference comprehension. *Cognitive Science* **44**, e12869.

91.

Kachel G, Hardecker DJ, Bohn M. 2021 Young children's developing ability to integrate gestural and emotional cues. *Journal of Experimental Child Psychology* **201**, 104984.

92.

Itakura S. 1996 An exploratory study of gaze-monitoring in nonhuman primates 1. *Japanese Psychological Research* **38**, 174–180.

93.

Bohn M, Call J, Tomasello M. 2016 Comprehension of iconic gestures by chimpanzees and human children. *Journal of Experimental Child Psychology* **142**, 1–17.

94.

Arkel J van, Woensdregt M, Dingemanse M, Blokpoel M. 2020 A simple repair mechanism can alleviate computational demands of pragmatic reasoning: Simulations and complexity analysis. In *Proceedings of the 24th conference on computational natural language learning*, pp. 177–194. Sl: Association for Computational Linguistics.

95.

Scott-Phillips T. 2014 *Speaking our minds: Why human communication is different, and how language evolved to make it special.* Macmillan International Higher Education.

96.

Melis A, Rossano F. 2022 Cooperation, coordination and communication in great apes. *Philosophical Transactions of the Royal Society B: Biological Sciences*

97.

Piantadosi ST, Tily H, Gibson E. 2012 The communicative function of ambiguity in language. *Cognition* **122**, 280–291.

98.

Hawkins RD, Franke M, Frank MC, Smith K, Griffiths TL, Goodman ND. 2021 From partners to populations: A hierarchical bayesian account of coordination and convention. *arXiv*

99.

Woensdregt M, Cummins C, Smith K. 2020 A computational model of the cultural co-evolution of language and mindreading. *Synthese*, 1–39.

100.

Fröhlich M, Sievers C, Townsend SW, Gruber T, Schaik CP van. 2019 Multimodal communication and language origins: Integrating gestures and vocalizations. *Biological Reviews* **94**, 1809–1829.

Acknowledgements

We thank Matthias Allritz, Mark Dingemanse and an anonymous reviewer for their helpful comments on previous versions of this paper.