A Functionalist Framework for Machine Learning in Animal Communication

Jack Terwilliger **Stephan P. Kaufhold** **Federico Rossano**
Department of Cognitive Science
University of California, San Diego
{jterwilliger, spkaufho, frossano}@ucsd.edu

Abstract

Recent breakthroughs in natural language processing inspire optimism that similar methods could decode animal communication systems. But machine learning approaches import assumptions from human language, which could undermine these efforts. In this proposal, we argue that non-human animal communication systems do not have self-contained distributional semantics, are largely non-referential, and function primarily to manipulate the behavior of others rather than exchange information. Not only do these assumptions constrain our ability to investigate signal semantics, but also risk confounding discoveries of signal syntax. To hedge against this possibility, we propose that machine learning efforts should adopt a functionalist framework. This foregrounds ecological and social contexts and the interactional contingencies that give signals their meaning. Our framework provides recommendations about how to account for these variables when building datasets.

1 Introduction

Large language models (T. Brown et al., 2020) can generate coherent sentences without relying on explicit linguistic domain knowledge or large amounts of manually supervised labels (Devlin et al., 2019). The success of LLMs has led some researchers to ask: what if we trained similar models on animal communication data? Could we begin to decode what whales, parrots, or bonobos are "saying" and even start a conversation?

Recently, research teams, initiatives, and prizes have begun pursuing these questions (Robinson et al., 2024; Rutz et al., 2023; Yovel & Rechavi, 2023; Sharma et al., 2024; Almeida et al., 2025). However, applying approaches from human language to animal communication imports several assumptions which could undermine these efforts. Although some animal communication systems display surface analogies with human language, they are not homologous, i.e., they do not stem from shared evolutionary origins (Scott-Phillips & Heintz, 2023). Here, we argue that non-human animal communication systems do not have self-contained distributional semantics, are largely non-referential, and function primarily to manipulate the behavior of others rather than exchange information with them. Without confronting these assumptions, researchers may be limited to answering questions about signal structure rather than questions about signal meaning. To overcome this, we propose that researchers adopt a functionalist framework by grounding signals in relevant behavioral responses beyond the communicative repertoire, social interactions, ecology, and development. With this in hand, we believe that machine learning efforts and domain experts will be better equipped to complement each other in the creation of new, large-scale, well-structured datasets.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Non-Human Animal Communication.

2 Three Questionable Assumptions from Human Communication

The possibility of learning useful representations of language from language itself is theoretically justified by the distributional hypothesis, which states linguistic items derive their meaning from the context of their use (Wittgenstein, 1953; Harris, 1954; Firth, 1957). Many NLP systems use purely linguistic context to embed items in a vector space encoding syntactic and semantic relationships (Mikolov et al., 2013). Humans also use distributional information to learn language (Romberg & Saffran, 2010). For example, congenitally blind individuals acquire color concepts that are structurally similar to those of sighted individuals via language (Lewis et al., 2019). However, humans learn from orders of magnitude less language input than language models (Warstadt & Bowman, 2022), because they also rely on rich extra-linguistic information, including social and embodied context. Nonetheless, the profound success of large language models is an important empirical finding that reflects how much distributional information in human language is present in text. However, when extending such approaches to animal communication, it is important to recognize the risk of implicitly importing assumptions drawn from human language, and especially those tied to textual training data of LLMs, that may not hold in these systems. Specifically, such efforts often presuppose (1) that meaning can be learned from signal sequences alone (self-contained distributional semantics), (2) that animal signals are referential (functioning like word-like labels for things), and (3) that the primary function of communication is information transfer rather than influencing others' behavior. We will briefly discuss these assumptions in relation to and outside human language.

Distributional information manifests through semiotic design features of language which allow for a rich matrix of relations between its elements. First and foremost, the syntax of language allows speakers to produce sentences in which they can embed similar words in similar grammatical constructions. But beyond the confines of a sentence, language is related across turns by discourse relevance. And, in its core niche, in conversation (Levinson, 2006), speakers' turns enact complementary social actions and language in subsequent turns can refer back to, reuse, and transform prior talk.

In contrast, animal communication systems lack these design features of language, which constrains the amount of self-contained distributional information available. Combinatorial syntax is rare and still debated (Bolhuis et al., 2018). To our knowledge, there is nothing like human discourse, and turn-taking in most animals is far simpler than in humans (Rossano, 2018), where innumerable conversational moves can be exchanged over durations of many hours. Moreover, unlike human language research, where abundant amounts of written texts and narratives provide training material, animal communication has no equivalent archives. LLMs are primarily trained on written text rather than on data of naturalistic human interactions, whereas animal signaling is fundamentally tied to moment-to-moment social interaction. This asymmetry poses challenges and the data distributions that enabled LLMs' success may not be available for animal communication systems.

But even large language models can fail to learn semantic information from distributional information if it involves embodied affordances (Jones et al., 2022), e.g., *he used his shirt to dry his feet* [afforded] vs. *he used his glasses to dry his feet* [non-afforded] (Glenberg & Robertson, 2000). A longstanding debate in cognitive science is whether the semantics of language results from being grounded in the world and in biological processes (Searle, 1980) or whether the context of is sufficient. Arguably, grounding may be achieved through multimodal models, which process information other than text (Lu et al., 2019). For example, visual language models jointly embed language and visual features (Radford et al., 2021), allowing models to perform tasks like captioning images, thus using language to refer to aspects of visual scenes. Moreover, recent approaches in animal communication have been training generative models on bioacoustic data Robinson et al. (2024) which embeds animal calls in a broader acoustic context, including sounds from the environment and those that result from their actions. It is tempting to think that we could derive a referential mapping between animal signals and events in the world by using such multimodal datasets.

However, whether animals understand reference is a debated topic in cognitive science. While domesticated animals can readily learn to comprehend labels that map onto objects and events (Kaminski et al., 2004; Bastos et al., 2024), such mappings are rare in endogenous communication systems. A famous example of such a system is in vervet monkeys (Seyfarth et al., 1980), which produce different alarm calls for different predators. Distributionally, these types of predators and the calls co-occur – they have high mutual information. But so do the associated differential behavioral responses. It is unknown whether vervet calls are analogous to a referential *eagle* label, i.e., functionally referential (Seyfarth et al., 1980), a performative *look up!*, or merely a predictive

cue (Wheeler & Fischer, 2012). We can be confident, however, that vervets are not using these alarm calls to talk about eagles in the richer, discourse-based sense.

The informational perspective, often framed by the Shannon-Weaver code model (Shannon, 1948; Weaver, 1963), has been and remains highly influential in the study of animal communication. This perspective views communication as a process of sending information, where signals are treated as containers carrying predefined messages from sender to recipient. However, such a code model minimizes or ignores the specific context and social relationship in which the signal occurs.

In contrast to the informational perspective, Dawkins & Krebs (1978, 1979) proposed that animal signals evolved to manipulate or influence the behavior of recipients to the sender's benefit. In functionalist views like the manipulation perspective, communication is best understood as a tool for influencing others and coordinating social interactions rather than exchanging information. Accordingly, signals are actions for achieving specific goals and consequences contingent on their social context. The meaning of a signal lies in its function or effect rather than in a consistent piece of information that could be decoded.

Below, we argue that a functionalist framework, treating signals as tools for influencing others within specific social and ecological contexts, can improve machine-learning approaches to animal communication. Importantly, it avoids dependence on the assumptions outlined above, making progress possible even when those are not met.

3 A Functionalist Framing

To hedge against issues the above assumptions may inject into training datasets, we propose machine learning research on animal communication adopt a functionalist framework. Grounded in ethology and comparative cognitive science, this framework can help guide how datasets are collected, ensuring they provide enough context to answer research questions about animal communication systems.

In contrast to the informational perspective, common in NLP, the functionalist perspective proposes that signals are best thought of as *tools for achieving specific goals that primarily benefit senders by manipulating recipients* (Figure S1B). The form and function of these signals may derive from many sources which affect the context needed to interpret them. But fundamentally, communicative signals are embedded in social interactions, which often involve instrumental, non-signaling behaviors.

To leverage the distributional hypothesis for animal communication in the functionalist framework, one must widen the scope of what counts as context. This can be clearly seen in the case of intentional communication. Intention underlies much of the theory in human communication. Townsend et al. (2017) operationalize intentional communication as signaling that is goal-oriented, contingent on the recipient's attention, and non-randomly eliciting a behavioral response conducive to achieving that goal. In such cases, a signal s derives its meaning from contexts in which signaler S has a goal g, a recipient is in attentional state a and makes a behavioral response r, which possibly affects g (Figure S1A). Therefore, for models to learn communicative signals, datasets may need to provide enough context from which the identities of signalers and recipients, their attentional states, and their goals/intentions can be inferred.

3.1 Signal Origin and Scope of Signaling Community

In humans, all of this is often made public through language, but in animals it can require observing a broad range of social and environmental context. The scope of this social context depends on the origin of the communicative behavior.

In evolution, signals often originate from instrumental acts (Dawkins & Krebs, 1978, 1979). Recipients remain sensitive to signals when there is mutual benefit, i.e., when, on average, perceiving them is advantageous. For example, an instrumental behavior such as biting, which originally had an immediate effect (injury) on a recipient, can ritualize into baring the teeth as a threat display. Because teeth baring reliably precedes biting, signalers can defend resources without physical contact, and recipients avoid injury. The process of phylogenetic ritualization occurs both within and across species.

In development, new communicative behaviors can emerge through ontogenetic ritualization (Tomasello, 2008). This process occurs when individuals repeatedly interact, anticipate one an-

other's instrumental actions for achieving a goal, and gradually abbreviate them into communicative signals. For example, bonobo infants initially raise their arms and tug on their mother's fur to get carried. But over time the gesture can be reduced to simply raising the arms as a carry request. In some cases, highly idiosyncratic signals can emerge, such as a bonobo initiating carries by spinning in place, a signal that was distinct to one mother-infant pair (Halina et al., 2013). For machine learning, this means that some signals have meaning only within a specific dyad's history. A general "translator" trained on population-level datasets may therefore miss or misinterpret signals whose function is rooted in individual interactional histories. Therefore, recognizing this limitation and knowing the origins of animal signals is crucial when curating datasets and evaluating model performance.

3.2 How Intentionality and Interaction Confound Compositionality

Intentional communication has some observable behavioral correlates, such as when an animal tries to elicit a response from another animal, it will repeat or reformulate signals until that intended response is elicited (or some other stopping condition has been met). For example, non-human primates will continue making *groom me* gestures until they are socially groomed (Byrne et al., 2017), dogs will continue to request food until they are fed (Worsley & O'Hara, 2018), and, in humans, if a conversational turn is not responded to, it will be pursued (Pomerantz, 1983; Stivers & Rossano, 2010). The response that stops signaling is referred to as an *apparently satisfactory outcome* (ASO) (Hobaiter & Byrne, 2014). ASOs are often used to infer the meaning of signals.

In humans, such response mobilization strategies are thought to influence the structure of language and its development (R. Brown, 1968; Goodwin, 2018; Du Bois, 2014; Terwilliger & Rossano, 2025). For animal communication research this means that, unless datasets specify who is signaling, whether the recipient responded (possibly in another modality), and the temporal structure of the exchange, analyses of signal structure risk being confounded by the dynamics of turn-taking (Figure S1C). For instance, if we observe the sequence *aab*, it could reflect a compositional syntax *or* it could be three separate communicative bouts, where a signaler repeats *a* twice and then tries *b* in pursuit of a response. The distinction hinges on timing: interactional silences, pauses, and bout boundaries all matter for interpreting structure. Without this information, models may conflate persistence strategies with grammatical composition. Prior work on the structure and timing of human versus non-human interaction highlights how central temporal dynamics are to communication (Rossano, 2019).

4 Recommendations for Dataset Design and Evaluation

We argued for the benefits of understanding animal communication through a functionalist framework, which treats signals not as isolated units but as social tools embedded within evolutionary dynamics, developmental trajectories, social relationships, and ecological contexts. This shift in perspective has concrete implications for how datasets are curated and how ML models are evaluated. Below, we highlight three priorities:

First, datasets should be longitudinal and multimodal (Rutz et al., 2023). Functional meaning often emerges only through repeated interaction and over developmental time, such as the example of ontogenetic ritualization in great ape gestures discussed above. Further, signals can change as individuals mature, as dominance relations shift, or as traditions shift (e.g., whale song dialects). Capturing such processes requires long-term recordings of identified individuals or groups, enriched with multimodal inputs (audio, video, movement, and potentially physiological states).

Second, datasets should include rich contextual annotations, which can be based on the longitudinal and multimodal raw data. Datasets should include the identity of signalers and recipients, correlates of their attentional states, behavioral outcomes following a signal (including apparently satisfactory outcomes), and relevant environmental context. This provides models with a record of interactions from which functional patterns can be learned.

Third, the success of models should be defined in terms of their predictive power for behavioral and social outcomes. For instance, can a model anticipate how a recipient will respond to a given signal in a given context? Does its interpretation shift appropriately with changes in social dynamics? Explicitly linking datasets and models to Tinbergen's (1963) four levels of analysis (mechanism, development, phylogeny, adaptation) helps clarify what kind of functional questions models are equipped to answer.

In sum, we believe significant progress can be made by grounding models in function and context, rather than projecting the structure of human language onto other species.

References

Almeida, C., Vail, C., Hernandez-Blick, C., Reiss, D., & Armstrong, K. (2025). Interspecies communication: Bridging the communication gap. In *Animals in translation workshop* (p. 9-17). Interspecies Internet. Retrieved from `https://www.interspecies.io/publication`

Bastos, A. P., Evenson, A., Wood, P. M., Houghton, Z. N., Naranjo, L., Smith, G. E., ... others (2024). How do soundboard-trained dogs respond to human button presses? an investigation into word comprehension. *PloS One*, *19*(8), e0307189.

Bolhuis, J. J., Beckers, G. J., Huybregts, M. A., Berwick, R. C., & Everaert, M. B. (2018). Meaningful syntactic structure in songbird vocalizations? *PLoS biology*, *16*(6), e2005157.

Brown, R. (1968). The development of wh questions in child speech. *Journal of verbal learning and verbal behavior*, *7*(2), 279–290.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... others (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Byrne, R. W., Cartmill, E., Genty, E., Graham, K. E., Hobaiter, C., & Tanner, J. (2017). Great ape gestures: intentional communication with a rich set of innate signals. *Animal cognition*, *20*(4), 755–769.

Dawkins, R., & Krebs, J. R. (1978). Animal signals: information or manipulation? In J. R. Krebs & N. B. Davies (Eds.), *Behavioural ecology: An evolutionary approach* (pp. 282–309). Oxford: Blackwell.

Dawkins, R., & Krebs, J. R. (1979). Arms races between and within species. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, *205*(1161), 489–511.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171–4186).

Du Bois, J. W. (2014). Towards a dialogic syntax. *Cognitive linguistics*, *25*(3), 359–410.

Firth, J. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, 10–32.

Glenberg, A. M., & Robertson, D. A. (2000). Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of memory and language*, *43*(3), 379–401.

Goodwin, C. (2018). *Co-operative action*. Cambridge University Press.

Halina, M., Rossano, F., & Tomasello, M. (2013). The ontogenetic ritualization of bonobo gestures. *Animal cognition*, *16*(4), 653–666.

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2-3), 146–162.

Hobaiter, C., & Byrne, R. W. (2014). The meanings of chimpanzee gestures. *Current Biology*, *24*(14), 1596–1600.

Jones, C. R., Chang, T. A., Coulson, S., Michaelov, J. A., Trott, S., & Bergen, B. (2022). Distributional semantics still can't account for affordances. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 44).

Kaminski, J., Call, J., & Fischer, J. (2004). Word learning in a domestic dog: evidence for" fast mapping". *Science*, *304*(5677), 1682–1683.

Levinson, S. C. (2006). On the human" interaction engine". In *Roots of human sociality* (pp. 39–69). Routledge.

Lewis, M., Zettersten, M., & Lupyan, G. (2019). Distributional semantics as a source of visual knowledge. *Proceedings of the National Academy of Sciences*, *116*(39), 19237–19238.

Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, *32*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, *26*.

Pomerantz, A. (1983). *Pursuing a response*. Cambridge University Press Cambridge, UK.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... others (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763).

Robinson, D., Miron, M., Hagiwara, M., Weck, B., Keen, S., Alizadeh, M., ... Pietquin, O. (2024). Naturelm-audio: an audio-language foundation model for bioacoustics. *arXiv preprint arXiv:2411.07186*.

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(6), 906–914.

Rossano, F. (2018). Social manipulation, turn-taking and cooperation in apes: Implications for the evolution of language-based interaction in humans. *Interaction Studies*, *19*(1-2), 151–166.

Rossano, F. (2019). The structure and timing of human versus primate social interaction. In P. Hagoort (Ed.), *Human language: From genes and brains to behavior* (pp. 201–219). MIT Press. doi: 10.7551/mitpress/10841.003.0019

Rutz, C., Bronstein, M., Raskin, A., Vernes, S. C., Zacarian, K., & Blasi, D. E. (2023). Using machine learning to decode animal communication. *Science*, *381*(6654), 152–155.

Scott-Phillips, T., & Heintz, C. (2023). Animal communication in linguistic and cognitive perspective. *Annual Review of Linguistics*, *9*(1), 93–111.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, *3*(3), 417–424.

Seyfarth, R. M., Cheney, D. L., & Marler, P. (1980). Monkey responses to three different alarm calls: evidence of predator classification and semantic communication. *Science*, *210*(4471), 801–803.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423.

Sharma, P., Gero, S., Payne, R., Gruber, D. F., Rus, D., Torralba, A., & Andreas, J. (2024). Contextual and combinatorial structure in sperm whale vocalisations. *Nature Communications*, *15*(1), 3617.

Stivers, T., & Rossano, F. (2010). Mobilizing response. *Research on Language and Social Interaction*, *43*(1), 3–31.

Terwilliger, J., & Rossano, F. (2025). Initiation asymmetry in the ontogenesis of social routines: In conversation, caregivers scaffold 1-year olds to respond, but 2-year olds initiate. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 47).

Tinbergen, N. (1963). On aims and methods of ethology. *Zeitschrift für tierpsychologie*, *20*(4), 410–433.

Tomasello, M. (2008). *Origins of human communication*. MIT press.

Townsend, S. W., Koski, S. E., Byrne, R. W., Slocombe, K. E., Bickel, B., Boeckle, M., ... others (2017). Exorcising g rice's ghost: An empirical approach to studying intentional communication in animals. *Biological Reviews*, *92*(3), 1427–1433.

Warstadt, A., & Bowman, S. R. (2022). What artificial neural networks can tell us about human language acquisition. In *Algebraic structures in natural language* (pp. 17–60). CRC Press.

Weaver, W. (1963). *The mathematical theory of communication*. University of Illinois Press.

Wheeler, B. C., & Fischer, J. (2012). Functionally referential signals: a promising paradigm whose time has passed. *Evolutionary Anthropology: Issues, News, and Reviews*, *21*(5), 195–205.

Wittgenstein, L. (1953). *Philosophical investigations*. Oxford: Blackwell.

Worsley, H. K., & O'Hara, S. J. (2018). Cross-species referential signalling events in domestic dogs (canis familiaris). *Animal Cognition*, *21*(4), 457–465.

Yovel, Y., & Rechavi, O. (2023). Ai and the doctor dolittle challenge. *Current Biology*, *33*(15), R783–R787.

Supplementary Material

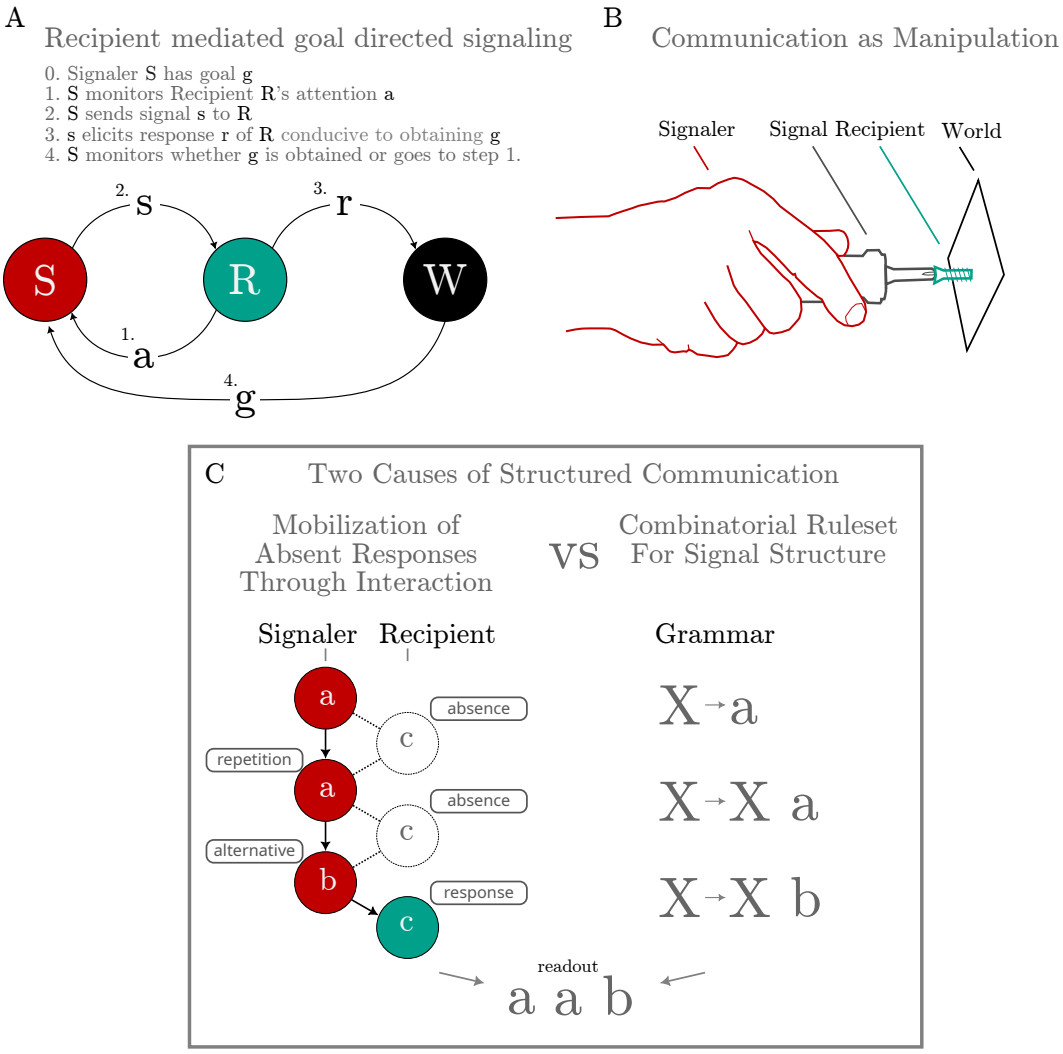

Figure S1: **(A) Recipient-mediated, goal-directed signaling.** A signaler *S* has a goal *g*, monitors a recipient *R*'s attentional state *a*, sends a signal *s* to *R*, and monitors the recipient's response *r*; signaling is repeated or reformulated until an apparently satisfactory outcome (ASO) is reached or a stopping condition occurs. **(B) Communication as manipulation.** Signals are actions selected by *S* to influence *R* so as to bring about desired changes in the world *W*, i.e., to achieve *g*. **(C) Two sources of structured sequences.** The same surface pattern (e.g., `a a b`) can arise from interactional dynamics (left), repetition after absent responses followed by an alternative, or from a genuine combinatorial ruleset for signal structure (right); disambiguation hinges on timing and bout boundaries.