

Extending rational models of communication from beliefs to actions

Theodore R. Summers¹, Robert D. Hawkins², Mark K. Ho¹, Thomas L. Griffiths^{1,2}

¹Department of Computer Science, Princeton University

²Department of Psychology, Princeton University

{sumers, rdhawkins, mho, tomg}@princeton.edu

Abstract

Speakers communicate to influence their partner’s *beliefs* and shape their *actions*. Belief- and action-based objectives have been explored independently in recent computational models, but it has been challenging to explicitly compare or integrate them. Indeed, we find that they are conflated in standard referential communication tasks. To distinguish these accounts, we introduce a new paradigm called *signaling bandits*, generalizing classic Lewis signaling games to a multi-armed bandit setting where all targets in the context have some relative value. We develop three speaker models: a *belief*-oriented speaker with a purely informative objective; an *action*-oriented speaker with an instrumental objective; and a *combined* speaker which integrates the two by inducing listener beliefs that generally lead to desirable actions. We then present a series of simulations demonstrating that grounding production choices in future listener actions results in relevance effects and flexible uses of nonliteral language. More broadly, our findings suggest that language games based on richer decision problems are a promising avenue for insight into rational communication.

Keywords: Communication, rational speech acts, multi-armed bandits, language games

Introduction

“Language is used for doing things.” (Clark, 1996, p. 3)

But *what* things? Broadly, accounts of communicative goals have been formulated in terms of listener *beliefs* and *actions*: “Alan is speaking with the aim of getting Barbara to understand him and to act on that understanding” (Clark, 1996, p. 11). But how do these aims relate? Is language primarily a tool for informing others, shaping their actions, or some combination of the two?

Classical accounts have emphasized *beliefs*, framing communication as information transfer between speaker and listener (Grice, 1975). This is reflected in the recent Rational Speech Acts (RSA) framework, which typically defines a speaker’s utility in terms of epistemic objectives like informativeness (Frank & Goodman, 2012; Kao, Wu, Bergen, & Goodman, 2014; Yoon, Tessler, Goodman, & Frank, 2020). A related perspective from the connectionist literature defines meaning in terms of effects on the listener’s latent representations (Elman, 2009). Yet beliefs themselves are an unsatisfying objective, as they are imperceptible and have no real-world consequences. Grice himself suggested informativeness should be generalized to “influencing or directing the actions of others” (Grice, 1975, p. 47).

Under an alternative *action*-oriented view, communication is an extension of an agent’s basic capability to interface

with the world, allowing a speaker to act indirectly through others. This has been explored in game-theoretic pragmatics (Qing & Franke, 2015; Benz & van Rooij, 2007; Franke & Jäger, 2016) and emergent communication (Lazaridou & Baroni, 2020). Practical natural language interfaces, including instruction-following (Tellex et al., 2011) and task-oriented dialogue systems (Chen, Liu, Yin, & Tang, 2017), typically learn direct mappings between commands and agent actions. While such imperative language is effective at inducing immediate actions, it cannot offer a full account of communication. Humans clearly employ more sophisticated strategies to “program” others (Lupyan & Bergen, 2016).

To reconcile these perspectives, we propose a unified computational model that integrates both action-based and belief-based objectives: communication *operates* by influencing intermediate beliefs, but its *objective* is to shape downstream actions in the world. Under this *combined* model, speakers reason about the decision problem a listener faces, how utterances may change their latent beliefs, and finally how those beliefs give rise to actions. They then choose utterances to induce belief states that maximize value in expectation over possible actions. Varying the scope of the decision problem allows reasoning at different time horizons, from “nudges” (Thaler & Sunstein, 2009) that encourage specific actions (e.g. getting someone to close a window by saying “It’s chilly in here”) to intervening on norms more broadly (Tomasello, 2016).

We first extend the Rational Speech Acts (RSA) framework (Frank & Goodman, 2012) to account for action-based objectives. We then introduce a new task paradigm, *signaling bandits*, which combines traditional signaling games (Lewis, 1969) with multi-armed bandits (Sutton & Barto, 2018). In Simulation 1, we show that standard reference games are a special case of the signaling bandit framework and conflate speaker objectives. In Simulation 2, we find that action-based speakers provide decision-relevant information, but employ hyperbolic and false utterances that improve decisions at the cost of distorting listener beliefs. Finally, Simulation 3 shows that when the Combined speaker reasons about a larger set of possible actions, it becomes more truthful and produces utterances that improve decision-making over the full distribution of contexts the listener could face. These results suggest that integrated reasoning over listener belief states and actions creates a pressure to transmit appropriately generalizable information.

Action-Grounded Speaker Models

Suppose a friend is about to step into a busy street. We have a brief moment to say something— but what should we say? The basic principle of *informativeness* suggests that we aim to reduce their uncertainty about the world as much as possible. Without a notion of *relevance* (Sperber & Wilson, 1986), however, it would be equally informative to mention a passing cloud or an incoming car. Recent accounts have formalized relevance as a *question under discussion* (QUD) that collapses the utility of the listener’s beliefs to a coarse-grained state, such as the current state of the road (e.g. Kao et al., 2014; Hawkins, Stuhlmüller, Degen, & Goodman, 2015). But our friend may benefit from a more outcome-grounded notion of relevance, projecting not to epistemic states but to the *decision problem* they face (Benz & van Rooij, 2007): intuitively, regardless of what they know, we want them to cross safely.

A simple way to ground relevance in the decision problem is to maximize the likelihood that the listener will take a specific action, effectively using that action as the QUD. This model, which we call the *action-oriented speaker*, would suggest something like “Cross after that car passes!” However, rather than choosing a premeditated action from our own perspective, we could instead give the listener information that allows them to act optimally from *their* perspective. This *combined* speaker aims to maximize the full expected utility of the listener’s actions under the induced beliefs. Such a speaker might say “Look both ways!” which encourages our friend to check for cars outside of our own field of view and (incidentally) generalizes across intersections.

Belief-oriented speaker. We formalize these models by first introducing the Rational Speech Acts framework (Frank & Goodman, 2012), which instantiates Gricean maxims (Grice, 1975) as recursive social inference. In this framework, speakers have knowledge of the world state w and choose between utterances u proportionally to their utility $U(u, w)$, where β_S is a soft-max parameter controlling speaker optimality:

$$P_S(u | w) \propto \exp\{\beta_S \cdot U(u, w)\} \quad (1)$$

Typically, this utility is defined in terms of the listener’s *beliefs*, their information gain about the state w :

$$U_{\text{Belief}}(u | w) = \log P_L(w | u) \quad (2)$$

This utility requires the speaker to reason about the listener’s expected beliefs after hearing the utterance:

$$P_L(w | u) \propto \delta_{\llbracket u \rrbracket(w)} P(w) \quad (3)$$

where $\delta_{\llbracket u \rrbracket(w)}$ represents the meaning of u , evaluating to one when utterance u is true of w and zero otherwise. This formulation optimizes for accurate listener beliefs, but lacks a notion of relevance.

Action-oriented speaker. Our second model extends the basic RSA framework to reason about actions a listener could take in the environment. Rather than adding an additional objective to the epistemic utility, we *ground* this objective

in a decision problem. Specifically, we re-formulate the listener as a reinforcement learning (RL) agent (Sutton & Barto, 2018) with a set of possible *actions* that may be taken, \mathcal{A} . At each point in time, a subset of those actions are available to the listener, which we call an *action context* $A \subseteq \mathcal{A}$. We assume the listener will choose actions to maximize a *reward* function, where the scalar reward value associated with each action is defined by the world state: $R : \mathcal{A} \times W \rightarrow \mathbb{R}$. We thus define the listener’s policy π_L in an action context A to be

$$\pi_L(a | u, A) \propto \exp\{\beta_L \cdot R_L(a, u)\} \quad (4)$$

where β_L is the softmax optimality and R_L is the listener’s expected reward for action $a \in A$ after hearing utterance u . We define expected reward with respect to their posterior beliefs about the likely state of the world:

$$R_L(a, u) = \sum_{w \in W} R(a, w) P_L(w | u) \quad (5)$$

In this work, we assume speakers are cooperative and have access to a ground-truth world state w . Action-based rational speakers reason about how their utterances will affect listener actions (Eqs. 3, 4, 5) and communicate to maximize the listener’s reward. We first describe a “pure” *action-oriented speaker*, which chooses the highest-reward action $a^* \in A$ and optimizes the probability of the listener taking that action:

$$a^* \triangleq \arg \max_{a \in A} R(a, w) \quad (6)$$

$$U_{\text{Action}}(u | A, a^*) = \log[\pi_L(a^* | u, A)] \quad (7)$$

Critically, rather than aiming for high-performing beliefs in general, this speaker only considers the listener’s beliefs insofar as they are relevant for producing the pre-selected action. Intuitively, this can be thought of as imperative language.

Combined speaker. Our final speaker unifies belief- and action-oriented objectives by optimizing over both, inducing beliefs which are likely to maximize rewards *in general*:

$$U_{\text{Combined}}(u | A, w) = \sum_{a \in A} \pi_L(a | u, A) R(a, w) \quad (8)$$

Effectively, this utility shifts the locus of decision-making to the listener: the combined speaker treats them as an independent agent and optimizes their beliefs, rather than choosing an action for them.¹ We now introduce an experimental setting to explore these speaker models.

Signaling Bandits

We define a new language game which enables study of the speaker models described above. In this setting, speakers cannot directly signal a unique correct action because all actions have some relative value. They must instead supply partial

¹Note our two action-oriented speakers can be recovered from a more general utility introducing an additional soft-max over the reward term $R(a, w)$, yielding the Action utility with $\beta \rightarrow \infty$ and the Combined utility with $\beta = 1$.

information to guide decision making. We first review the structure of Lewis signaling games and note their limitations. We then describe multi-armed bandits, a setting studied in reinforcement learning. Finally, we combine the two to produce a new game, *signaling bandits*.

Lewis Signaling Games

Lewis signaling games are two-player collaborative settings with a speaker and a listener (Lewis, 1969). Following the notation introduced previously, a signaling game is defined by a world state w , action context available to the listener, A , and utterances available to the speaker, \mathcal{U} . There is one action $a^* \in A$ with a positive reward; other actions have zero reward. The world state w implies the correct action a^* . The speaker knows w but the listener does not. During gameplay, the speaker chooses an utterance $u \in \mathcal{U}$ and sends it to the listener. The listener updates their beliefs, $P_L(w | u)$, and uses the posterior to choose an action, $\pi_L(a | u, A)$.

Signaling games formalize the coordination problem underlying communication (Krahmer & Van Deemter, 2012; Frank & Goodman, 2012). However, the interplay of beliefs, actions, and rewards is highly constrained. The state of the world w is synonymous with the correct action a^* , and players are indifferent over other actions. It is thus impossible to discriminate the three speaker objectives defined above (Eq. 2, 7, 8).² For a richer decision-making setting, we turn to multi-armed bandits.

Multi-Armed Bandits

A multi-armed bandit is a single-player sequential game. In each round, the player takes an action and receives a scalar reward (Sutton & Barto, 2018). Players seek to maximize their rewards, but are initially ignorant of the reward structure. Over multiple rounds, they must balance exploration (choosing a new action to learn its reward) with exploitation (choosing the most valuable known action). Because payoffs are scalar, decisions are more nuanced than Lewis games.

Contextual bandits extend this to study learning via abstract information. Actions are now characterized by features, and rewards are defined with respect to these features. Formally, a feature function ϕ describes actions: $\phi : A \rightarrow \mathbb{R}^K$. Rewards are then defined as a function of these features: $R : \phi(a) \rightarrow \mathbb{R}$. Thus, rather than learn about the reward of a specific action, players can learn about the reward of a *feature* which applies to many actions. For example, an animal might learn that ripe yellow bananas are high-reward, while rotten brown bananas are low-reward. Associating the payoff with the color (a feature) rather than the banana (a specific action) allows knowledge to transfer to new settings (the next banana). Contextual bandits have been studied extensively in reinforcement learning and to a lesser degree for emergent communication (Donaldson, Lachmann, & Bergstrom, 2007). Yet to our knowledge, they have not been used to

²But see Qing and Franke (2015) for evidence in favor of action-oriented speakers.

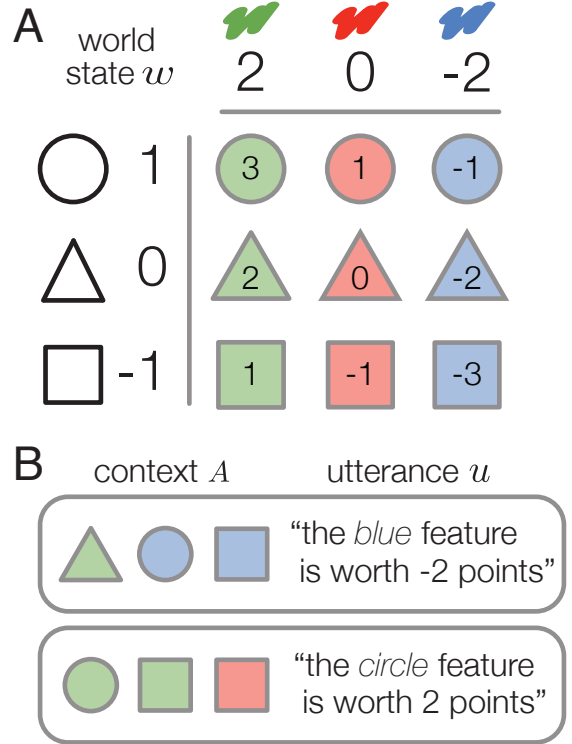


Figure 1: Signaling bandits combine Lewis signaling games with multi-armed bandits. (A) The world w is defined by correspondences between features and rewards (table margins), which combine additively to create possible actions \mathcal{A} (table contents). (B) Two contexts A with example utterances. The first utterance is true and useful (encouraging the listener to avoid the blue objects). The second is false (circles are worth 1). It is locally useful (encouraging the listener to choose the circle) but will lead the listener astray in other contexts.

study communication with an existing language. In the next section, we introduce a two-player version of this game.

Signaling Bandits

We combine the communication of Lewis games with the reward structure of contextual bandits to create a new game, *signaling bandits*. Unlike Lewis games, speakers no longer communicate concrete information (which action is correct). Instead, they communicate abstract information (how much features are worth). We now describe basic gameplay.

As in Lewis games, signaling bandits are two-player games with a speaker and a listener. Each game is defined by a world state w , a set of all possible actions \mathcal{A} and a set of speaker utterances \mathcal{U} . In each round, the listener faces an action context $A \subseteq \mathcal{A}$. However, *unlike* Lewis games, there is no single "correct" action. Instead, as in contextual bandits, each action has a scalar reward defined by the world state w . Here, we assume features are indicator variables over actions:

$$\phi : A \rightarrow \{0, 1\}^K \quad (9)$$

Rewards R are linear over these features, parameterized by w :

$$R(a, w) = w^\top \phi(a). \quad (10)$$

Concretely, this means the world state w is a vector encoding the reward associated with each feature. A listener with full world knowledge (every element of w) can calculate exact rewards using Eq. 10 and thus select the optimal action in any context. Fig. 1A depicts this visually: w defines the value of individual shapes and colors (table margins), which in turn determines the reward for each possible action in \mathcal{A} (table contents). The value of an individual feature (one element of w) thus constitutes *partial* knowledge about the world.

The speaker helps the listener by providing such partial knowledge. Formally, \mathcal{U} is a set of tuples of the form $\langle \mathbb{1}_K, \mathbb{R} \rangle$ which specify a given feature and scalar value. As shown in Fig. 1, these are messages like $\langle \text{Blue}, -2 \rangle$ or $\langle \text{Circle}, 2 \rangle$. Speakers choose utterances and send them to the listener. The (literal) listener updates their beliefs by setting the corresponding feature to the transmitted scalar value:

$$p_L(w_L | u) = p_L(w_L | u_{\mathbb{1}_K} = u_{\mathbb{R}}) \quad (11)$$

and then chooses an action from the available set A according to their posterior belief over rewards (Eqs. 4 and 5).

Transmitting partial knowledge and constructing different contexts $A \subseteq \mathcal{A}$ introduces several important dynamics. First, it induces relevance effects, since different knowledge will be useful in different contexts. For example, in Fig. 1B, $\langle \text{Green}, 2 \rangle$ would improve decision making in both contexts while $\langle \text{Blue}, -2 \rangle$ would only be relevant for the top one. Second, it accommodates nonliteral language naturally, as false beliefs can yield good decisions. In the bottom context of Fig. 1B, a false message $\langle \text{Circle}, 2 \rangle$ maximizes the probability of the listener choosing the optimal action. Finally, it allows us to explore *generalization*: whether a listener’s beliefs facilitate good decision making over other action contexts constructed from the same world. A bias towards communicating generalized information is implicated in cultural learning (Csibra & Gergely, 2009; Tessler & Goodman, 2019); thus, modeling the dynamics of speaker objectives and resulting generalization performance is of significant theoretical interest.

Signaling bandits creates clean distinctions between listener beliefs $P_L(w | u)$, the optimal action a^* , and the value of individual actions, $R(a, w)$. This allows for meaningful differences between speaker models (Eqs. 2, 7, 8). In the following section, we use simulations to illustrate this. We return to extensions beyond basic gameplay in the General Discussion.

Simulations

We perform three simulations within the signaling bandits framework. We first describe the general procedure and metrics used to measure speaker behaviors. For all simulations, we set $\beta_L = 3, \beta_{\text{Belief}} = 3, \beta_{\text{Action}} = 3$, and $\beta_{\text{Combined}} = 2$. Speaker optimality does not affect the qualitative results; we analyze optimal speakers ($\beta \rightarrow \infty$) at the end of this section. We assume listeners have uniform priors over feature rewards, and allow speakers to send only a single message.

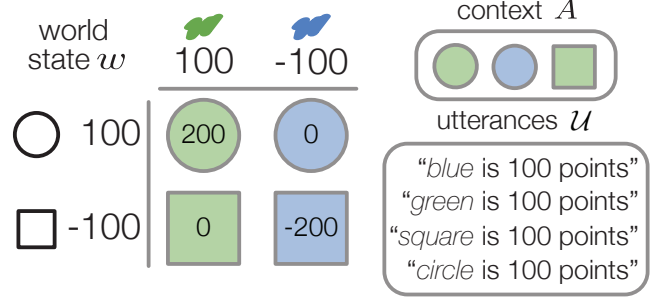


Figure 2: Simulation 1 constructs a traditional Lewis signaling game within the signaling bandits framework. Note that two of the possible utterances are false.

Speaker	$P_S(\text{truthful})$	$\pi_L(a^*)$	$R_S(A)$	$R_S(\mathcal{A})$
Belief	1.00	.500	100	-
Action	1.00	.500	100	-
Combined	1.00	.500	100	-

Table 1: Simulation 1 results. Reference games align all three speakers’ objectives and so cannot disambiguate them. Generalization, $R_S(\mathcal{A})$, does not apply to this setting.

Procedure. For each simulation, we define a world state w and set of allowable utterances \mathcal{U} . For an action context A , we first compute each speaker’s distribution over utterances $u \in \mathcal{U}$ as defined by Eqs. 2, 7, and 8. We then compute the listener’s resulting distribution over actions (Eq. 4).

Evaluation Metrics. We use four metrics to summarize speaker behavior. First, their probability of choosing a true utterance, $P(\text{truthful})$. Second, the probability of the listener choosing the optimal action in A (Eq. 4), which we write as $\pi_L(a^*)$ for brevity. Third, the expected reward on the action context A (Eq. 8), which we write as $R_S(A)$. Finally, we want to know whether speakers are over-optimizing for a particular context. To evaluate this, we calculate the *expected generalization*, which indicates whether the listener’s resulting belief state yields good performance on other contexts drawn from \mathcal{A} . Formally, we compute the expected reward of the speaker’s utterance u across all possible contexts:

$$R_S(u, w, \mathcal{A}) = \sum_{A \in [\mathcal{A}]^3} R_S(u, w, A) P(A) \quad (12)$$

where $P(A)$ is the probability of an action context A ; here, we assume a uniform distribution over all contexts of size 3. Again, we shorten this to $R_S(\mathcal{A})$ for clarity. If local performance substantially exceeds generalization, $R_S(A) \gg R_S(\mathcal{A})$, we say the speaker generalizes poorly: it is optimizing local decision-making by providing false or less broadly useful information.

Simulation 1: Reproducing reference games

Our first simulation constructs a Lewis signaling game as a special case of our more general class of signaling bandits (Fig. 2). We show that this case cannot distinguish between our models, motivating Simulations 2 and 3.

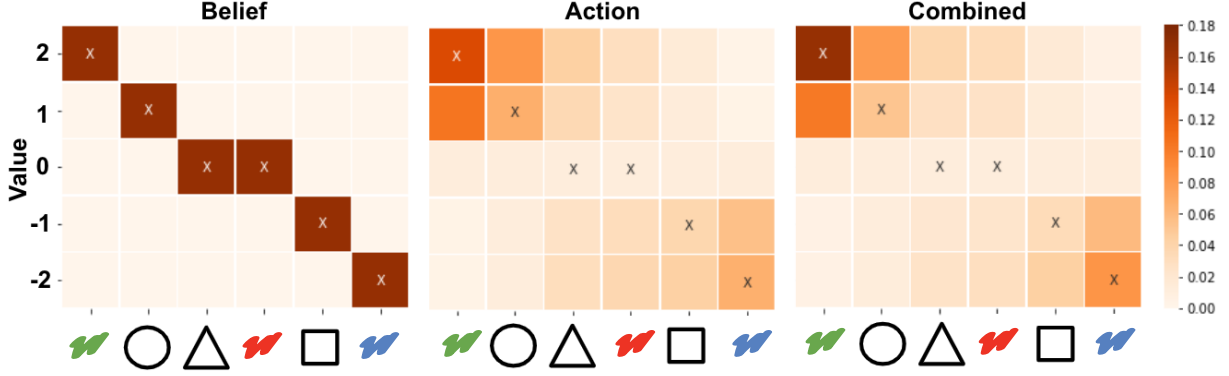


Figure 3: Simulation 2 results. Shading indicates probability of speaker choosing that utterance, averaged over all 84 3-action contexts. X’s indicate true utterances. Left: Belief speakers choose true utterances at random. Center/Right: Action and Combined speakers focus on decision-relevant features and exaggerate to improve listener decisions.

Speaker	$P_S(\text{truthful})$	$\pi_L(a^*)$	$R_S(A)$	$R_S(\mathcal{A})$
Belief	1.00	.499	.539	.539
Action	.330	.772	1.18	.486
Combined	.360	.742	1.28	.522

Table 2: Simulation 2 results. Action and Combined frequently send false messages. They obtain high performance on the local context, $R_S(A)$, but generalize poorly, $R_S(\mathcal{A})$.

Speaker	$P_S(\text{truthful})$	$\pi_L(a^*)$	$R_S(A)$	$R_S(\mathcal{A})$
Belief	1.00	.499	-	.539
Action	.440	.566	-	.748
Combined	.534	.627	-	.949

Table 3: Simulation 3 results. When communicating about a larger action context, Action and Combined speakers become more truthful and generalization improves dramatically.

Setup. To construct a Lewis game with a single target, we define w and construct a context A containing one action with arbitrarily high reward (the target) and two with zero reward (the distractors). We restrict utterances \mathcal{U} to positive messages (corresponding to possible referential labels).

Results and Discussion. Results are summarized in Table 1. All three speaker objectives are aligned, and so we find that they behave identically: they choose between the two literally true messages $\langle \text{Green}, 100 \rangle$ or $\langle \text{Circle}, 100 \rangle$. No model has any reason to prefer a false utterance, or to prefer one true utterance over the other.

Simulation 2: Divergent speaker behaviors

We next explore how different speaker models may diverge for other tasks in our signaling bandit paradigm (Fig. 1).

Setup. We consider the world state depicted in Fig. 1. We set \mathcal{U} to all feature-value tuples and evaluate each speaker’s behavior across all $\binom{9}{3} = 84$ possible contexts of 3 actions.

Results and Discussion. We plot each speaker’s probability over individual utterances in Fig. 3, and summarize the results in Table 2. All metrics are averaged across the 84 contexts. First, we observe that Belief speakers choose a true utterance at random, regardless of the action context. This yields relatively poor performance locally but perfect generalization ($R_S(A) = R_S(\mathcal{A}) = .539$). In contrast, both Action and Combined speakers lie frequently ($P_S(\text{truthful}) < .5$). Action speakers tailor their utterances to induce the single optimal action in each set. As a result, they exaggerate whichever

feature values align with the best action in the immediate context. This strategy succeeds locally: they induce the most-optimal action a majority of the time ($\pi_L(a^*) = .772$) and obtain more than twice the reward obtained by the Belief speaker ($R_S(A) = 1.18$). However, the resulting distortion in listener beliefs means they generalize poorly ($R_S(\mathcal{A}) = .486$). Finally, Combined speakers achieve a middle ground. They obtain the best outcome less frequently than Action speakers ($\pi_L(a^*) = .742$), but higher reward locally ($R_S(A) = 1.28$) and generally ($R_S(\mathcal{A}) = .522$). Sensitivity to the rewards of all three actions leads them to distort beliefs less than Action speakers. We visualize the divergence between Action and Combined speakers in Fig. 4 in a single action context to better understand these differences.

Simulation 3: Expanding speaker context

Simulation 2 showed that Action and Combined speakers can be myopic: they produce messages to induce locally-optimal actions at the cost of generalization. Simulation 3 explores how this changes when they optimize over the entire action space \mathcal{A} . We find that both Action and Combined speakers become more truthful and generalize better.

Setup. We use the same world as Simulation 2 (shown in Fig. 1). We first construct a single “global” action context of all 9 actions: $A = \mathcal{A}$. We compute each speaker’s distribution over utterances for this 9-action context, then evaluate generalization over 3-action contexts, $R_S(\mathcal{A})$. Because there is no “local” context, we do not compute $R_S(A)$.

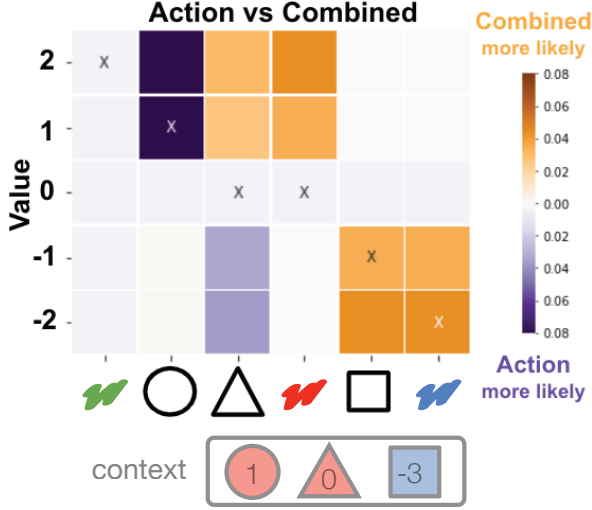


Figure 4: Comparing Action and Combined in one action context (Simulation 2). Shading indicates difference in utterance probabilities between speakers. The Action speaker maximizes the listener’s probability of choosing the red circle: $\pi_L(a^*) = .737, R_S(A) = .456$. The Combined speaker achieves higher reward by avoiding the blue square: $\pi_L(a^*) = .627, R_S(A) = .482$. Neither sends messages about green, demonstrating relevance effects.

Results and Discussion. Results are summarized in Table 3. The Belief speaker is unchanged. Both action-oriented speakers are more truthful and generalize better than Simulation 2. However, the Action speaker fixates on the *single best* option (the green circle). It sends false messages which exaggerate its value, e.g. $\langle \text{Circle}, 2 \rangle$, or discourage alternatives, e.g. $\langle \text{Red}, -2 \rangle$. As a result, it fares poorly whenever a green circle is not present. This illustrates the brittleness of optimizing to obtain a specific action. Because the Combined speaker optimizes in expectation over all actions, it is more likely to send true messages about extreme values, e.g. $\langle \text{Green}, 2 \rangle$ or $\langle \text{Blue}, -2 \rangle$. It obtains both higher rewards $R_S(\mathcal{A})$ and the optimal action $\pi_L(a^*)$ more frequently.

Effects of speaker optimality

While we fixed the speaker optimality parameter β_S throughout our simulations, it may interact in important ways with our model comparison. First, our Belief speaker is insensitive to β : since we assumed that the listener’s prior over w is uniform, all true utterances are equally valuable. By contrast, the Action and Combined speakers are sensitive to β in different ways, since they take the soft-max over different quantities (Action over log-probabilities and Combined over expected utilities). At any given β , Combined dominated Action on all metrics, so we tuned β to equate them ($\beta_{\text{Action}} = 3, \beta_{\text{Combined}} = 2$). At the same time, it is informative to explore their asymptotic behavior as $\beta \rightarrow \infty$, which is summarized in Table 4. In Simulation 2, we find that Action and Combined still send false messages and generalize poorly in the limit. Yet when multiple messages are equally likely to

Context	Speaker	$P_S(\text{truthful})$	$\pi_L(a^*)$	$R_S(A)$	$R_S(\mathcal{A})$
Both	Belief	1.00	.499	.539	.539
Local	Action	.458	.942	1.63	.690
Local	Combined	.488	.942	1.67	.710
Global	Action	.500	.624	-	.958
Global	Combined	1.00	.731	-	1.28

Table 4: Speaker behavior as $\beta \rightarrow \infty$. *Local* context corresponds to Simulation 2, and *Global* context to Simulation 3.

induce an optimal action, the Combined speaker converges to preferring truthful ones while the Action speaker is indifferent. This difference results in dramatically better performance in Simulation 3: the Combined speaker converges on producing the utterance $\langle \text{Green}, 2 \rangle$, while the optimal Action speaker remains ambivalent between $\langle \text{Green}, 2 \rangle$ and $\langle \text{Circle}, 2 \rangle$. In sum, the Combined speaker’s sensitivity to underlying reward structure leads it to consistently produce more truthful and generalizable utterances across contexts and in the limit of optimality.

General Discussion

Humans communicate to influence one another’s beliefs *and* actions. Here, we explored different ways speakers can reason about these objectives. We introduced two action-oriented speaker models which optimize for a downstream *decision problem*, grounding relevance in the listener’s actual decision context (Roberts, 2012). Critically, we proposed that rational “Combined” speakers should consider both beliefs and actions: they should communicate to induce belief states that are likely to produce high-value actions. To distinguish speaker models, we introduced a new communication game, signaling bandits. Signaling bandits generalizes Lewis signaling games to multi-armed bandits, formalizing communication in richer decision settings. Simulations show that the Combined speaker prefers *generalizable* information that is likely to produce high-value actions across a distribution of possible future contexts. This finding raises intriguing connections to belief-oriented accounts of generics (e.g. “Birds fly”; Tessler & Goodman, 2019) as well as biases towards generalizable examples in non-linguistic pedagogy (Csibra & Gergely, 2009; Tomasello, 2016). We are thus optimistic that such speaker models may provide a bridge from communicative principles to social learning more broadly.

This work represents a small step towards a deeper exploration of action-grounded models of rational communication. First, human experiments are needed to validate our simulations. Second, we considered only literal listeners in collaborative settings. Pragmatic listeners may reason about a speaker’s objectives and knowledge, as well as the action context the speaker considered (Goodman & Stuhlmüller, 2013). Finally, we explored only single-round gameplay; iterated games would allow for richer interactions. Speakers could observe listener actions and infer their beliefs via inverse reinforcement learning (Ng & Russell, 2000). A single message

followed by learner actions in multiple contexts would force speakers to optimize for a distribution over contexts. This would make generalization an objective rather than an incidental effect, as in optimal reward design (Singh, Lewis, & Barto, 2009). Listeners could learn both socially (via speaker messages) and individually (via their own actions). We hope we have successfully signaled the high value of research in this paradigm!

Acknowledgements

We thank our anonymous reviewers for their thoughtful feedback. This work was supported by NSF grant #1545126 and John Templeton Foundation grant #61454 to TLG and NSF grant #1911835 to RDH.

Code for simulations available at:
<https://github.com/tsumers/signaling-bandits>

References

- Benz, A., & van Rooij, R. (2007). Optimal assertions, and what they implicate. A uniform game theoretic approach. *Topoi*, 26(1), 63–78.
- Chen, H., Liu, X., Yin, D., & Tang, J. (2017). A survey on dialogue systems: Recent advances and new frontiers. *ACM SIGKDD Explorations Newsletter*, 19(2), 25–35.
- Clark, H. H. (1996). *Using Language*. Cambridge University Press.
- Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, 13(4), 148–153.
- Donaldson, M. C., Lachmann, M., & Bergstrom, C. T. (2007). The evolution of functionally referential meaning in a structured world. *Journal of Theoretical Biology*, 246(2), 225–233.
- Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science*, 33(4), 547–582.
- Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336(6084), 998–998.
- Franke, M., & Jäger, G. (2016). Probabilistic pragmatics, or why Bayes’ rule is probably important for pragmatics. *Zeitschrift für Sprachwissenschaft*, 35(1), 3–44.
- Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, 5(1), 173–184.
- Grice, H. P. (1975). Logic and conversation. In *Syntax and Semantics: Vol. 3: Speech Acts* (p. 41–58). New York: Academic Press.
- Hawkins, R. D., Stuhlmüller, A., Degen, J., & Goodman, N. D. (2015). Why do you ask? Good questions provoke informative answers. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society*.
- Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, 111(33), 12002–12007.
- Krahmer, E., & Van Deemter, K. (2012). Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1), 173–218.
- Lazaridou, A., & Baroni, M. (2020). Emergent multi-agent communication in the deep learning era. *arXiv preprint arXiv:2006.02419*.
- Lewis, D. (1969). *Convention: A Philosophical Study*. John Wiley & Sons.
- Lupyan, G., & Bergen, B. (2016). How language programs the mind. *Topics in Cognitive Science*, 8(2), 408–424.
- Ng, A. Y., & Russell, S. J. (2000). Algorithms for inverse reinforcement learning. In *Proceedings of the 17th International Conference on Machine Learning*.
- Qing, C., & Franke, M. (2015). Variations on a Bayesian theme: Comparing Bayesian models of referential reasoning. In *Bayesian Natural Language Semantics and Pragmatics* (pp. 201–220). Springer.
- Roberts, C. (2012). Information structure: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, 5, 6–1.
- Singh, S., Lewis, R. L., & Barto, A. G. (2009). Where do rewards come from? In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*.
- Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition* (Vol. 142). Harvard University Press Cambridge, MA.
- Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT press.
- Tellex, S., Kollar, T., Dickerson, S., Walter, M. R., Banerjee, A. G., Teller, S., & Roy, N. (2011). Understanding natural language commands for robotic navigation and mobile manipulation. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence*.
- Tessler, M. H., & Goodman, N. D. (2019). The language of generalization. *Psychological Review*, 126(3), 395.
- Thaler, R. H., & Sunstein, C. R. (2009). *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Penguin.
- Tomasello, M. (2016). Cultural learning redux. *Child Development*, 87(3), 643–653.
- Yoon, E. J., Tessler, M. H., Goodman, N. D., & Frank, M. C. (2020). Polite speech emerges from competing social goals. *Open Mind*, 4, 71–87.