Selective Truths: How Speakers Mislead and Listeners Miss It?

Ke Fang[1], **Ruijia (Hannah) Guan**[2], **Charley M. Wu**[4,5], **Michael Franke**[6], **Robert D. Hawkins**[3]

fangke@stanford.edu — rjguan@stanford.edu — rdhawkins@stanford.edu
charley.wu@tu-darmstadt.de — mchfranke@gmail.com

[1] Department of Psychology, Stanford University [2] Symbolic Systems Program, Stanford University
[3] Department of Linguistics, Stanford University [4] Center for Cognitive Science, TU Darmstadt
[5] Hessian.AI, TU Darmstadt [6] Department of Linguistics, University of Tübingen

Abstract

How do speakers craft misleading messages while remaining literally truthful? Can listeners resist such manipulation? We propose that while speakers rely on Theory of Mind to convey strategically underinformative messages, listeners need the same capacity—reasoning about speakers' goals—to avoid being misled. We formalize this by extending the Rational Speech Act framework to speakers with persuasive goals. The model predicts that (1) persuasive speakers rationally exploit underinformativity to frame information favorably, but (2) listeners' ability to correct for such bias depends on whether they actively reason about speaker goals. Experiment 1 confirms that human speakers systematically shift toward underinformativity when the evidence conflicts with their persuasive goals. Experiment 2 reveals an asymmetry: despite knowing that speakers may have persuasive motives, listeners exhibit low sensitivity to underinformativity. These results suggest that recursive social reasoning in strategic communication may be asymmetric—speakers exploit pragmatic inference, but listeners do not fully reciprocate with epistemic vigilance.

Keywords: epistemic vigilance; pragmatics; Rational Speech Act; persuasion; underinformativity; Theory of Mind

Introduction

When only 2 of 10 patients improved in a clinical trial, a pharmaceutical representative may report that "some patients showed improvement." A scientist with the same data might instead say that "most patients showed no improvement." Both statements are literally true, yet they invite very different inferences about the effectiveness (Rogers et al., 2017). This is *selective truth-telling*: using underinformative but truthful language to shape beliefs while maintaining plausible deniability. Selective truths pose a puzzle for belief revision. Credulous listeners who take statements at face value will be misled (Kamenica & Gentzkow, 2011). But skeptical listeners who discard everything lose the benefits of communication altogether (Crawford & Sobel, 1982). How can listeners learn from potentially biased speakers without being deceived?

We propose that selective truths provide a subtle signal of speaker's bias, but one that can only be decoded through Theory of Mind (ToM). Selective truths differ from outright lies in a crucial way: speakers never say anything false. Detecting lies often requires access to ground truth or prior knowledge about the speaker. In selective truth-telling, however, the speaker's choice to be underinformative can itself be informative. When a speaker says "some" instead of giving precise numbers, listeners can ask: why did they choose to be vague? *Underinformativity* suggests that the speaker may have something to hide, if an informative speaker would have been more precise. Thus, listeners need not simply accept or reject the message wholesale. Instead, by reasoning about how truth has been selectively framed, they may use underinformativity as a cue to the speaker's goals.

This connects to broader work on *epistemic vigilance*—the suite of cognitive mechanisms that allow people to critically evaluate communicated information rather than blindly accepting it (Sperber et al., 2010). Prior accounts have emphasized tracking source reliability across repeated interactions (Corriveau et al., 2009), but this cannot explain vigilance toward novel sources or single utterances. Our account highlights a different kind of vigilance, one that operates on the content or form of an utterance rather than on a speaker's history or on independently verified ground truth.

Recent computational work on deception suggests that recursive ToM may provide the mechanism for such vigilance. Oey et al. (2023) showed that people design and detect outright lies through recursive social reasoning: speakers craft lies by inferring what receivers will detect, while receivers detect lies by inferring how speakers would lie. Critically, Oey and Vul (2024) demonstrated that listeners can even recover accurate beliefs from known lies when they are given information about speakers' directional motives and lying costs.

Following this intuition, we formalize selective truth-telling using the Rational Speech Act (RSA) framework (Degen, 2023; Goodman & Frank, 2016), extending it to contexts where speakers may be informative or persuasive. Our model shows that (1) persuasive speakers rationally exploit underinformativity to frame information favorably while remaining truthful, corroborating prior work on utterance choice in argumentative contexts (e.g., Macuch Silva et al., 2024), and that (2) vigilant listeners who reason about speaker goals can recover from biased framing, using recursive reasoning to undo the inferences that speakers exploit. We then report two behavioral experiments testing whether people produce strategic underinformativity and whether listeners are sensitive to it. Together, our results suggest that selective truth-telling and epistemic vigilance are complementary consequences of recursive social reasoning in strategic communication [1].

[1] Materials and data are available here. Experiments were coded using `jsPsych` (De Leeuw et al., 2023). AI tools were used.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

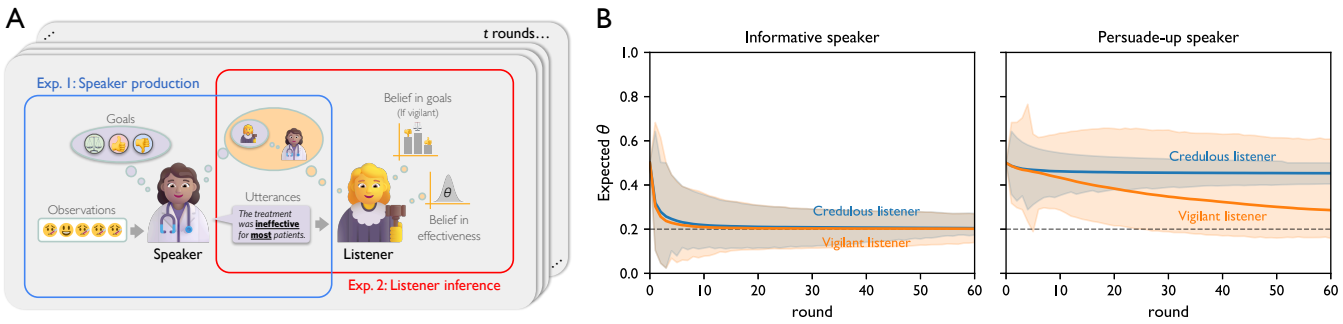

Figure 1: (A) Paradigm and experiments overview. (B) Simulated mean posterior belief over θ for **credulous vs vigilant listeners** communicating with **informative vs persuade-up speakers** for 60 rounds, across 5,000 simulated runs at true $\theta = 0.2$.

Computational Model

Formalization

Task Setting. A speaker observes clinical trial outcomes and describes them to a listener who aims to learn the treatment's effectiveness (Figure 1A). Each of n patients receives a treatment with unknown effectiveness rate $\theta \in [0,1]$, which determines the probability of improvement. The speaker observes the outcomes—say, k of n patients improved—and must communicate this observation using natural language. Rather than stating precise numbers, speakers produce utterances combining a quantifier $q \in \{no, some, most, all\}$ with a predicate $p \in \{effective, ineffective\}$:

"The treatment was [effective/ineffective] for [Q] patients."

Quantifiers have standard "at least" semantics: *some* is true if ≥ 1; *most* if $> n/2$; *all* if $= n$; and *no* if $= 0$. For utterances with the predicate *effective*, truth is evaluated with respect to the number of improved patients, k; for utterances with *ineffective*, truth is evaluated with respect to the number of non-improved patients, $n-k$. Crucially, multiple utterances can be true of the same observation. If 3 of 5 patients improved, both "effective for *some* patients" and "effective for *most* patients" are literally true, but they differ in informativeness.

Speaker Model. While a *literal speaker* S_0 selects uniformly among true utterances, a *pragmatic speaker* S_1 selects utterances to maximize a goal-dependent utility:

$$P_{S_1}(u \mid O, \psi) \propto \exp\left[\alpha \cdot V(u; O, \psi)\right] \cdot \mathbb{I}[\text{True}(u, O)] \quad (1)$$

where O denotes the observed outcomes, α controls speaker optimality, and $\mathbb{I}[\cdot]$ enforces literal truth.

The utility V depends on the speaker's goal ψ, for which we consider three types. An informative speaker ($\psi = \text{inf}$) maximizes $V_{\text{inf}}(u; O) = \log P_{L_0}(O \mid u)$, favoring utterances that help a literal listener recover the true observation. A persuade-up speaker ($\psi = \text{pers}^+$) maximizes $V_{\text{pers}^+}(u) = \mathbb{E}_{L_0}[\theta \mid u]$, favoring utterances that increase the listener's belief about θ. A persuade-down speaker ($\psi = \text{pers}^-$) maximizes $V_{\text{pers}^-}(u) = 1 - \mathbb{E}_{L_0}[\theta \mid u]$, favoring utterances that decrease the listener's belief about θ. This captures selective truth-telling: persuasive speakers choose among literally true utterances to bias the listener's inference.

Listener Model. A credulous listener assumes that the speaker is informative and interprets utterances accordingly, as in standard RSA. A vigilant listener instead maintains uncertainty over the speaker's goal ψ and jointly infers both the treatment effectiveness θ and the speaker's goal:

$$P_{L_1}(\theta, \psi \mid u) \propto P(\theta) \cdot P(\psi) \cdot P_{S_1}(u \mid \theta, \psi). \quad (2)$$

This captures epistemic vigilance by having the listener reason about *why* the speaker chose a particular utterance. If an informative speaker would have been more precise, deliberate underinformativity becomes evidence that the speaker may have a persuasive goal.

Model Predictions and Simulation

Building on this formalization, we describe the model's qualitative predictions for one-shot speaker production and listener inference, then examine how these patterns unfold over repeated interactions in simulation.

Strategic Underinformativity of Speakers. Informative and persuasive speakers produce distinct utterance patterns. Consider an observation in which 2/5 patients improve (Figure 2, Column 2)—weak evidence for the treatment. An informative speaker prefers the most specific true description: "ineffective for *most*." A persuade-down speaker also prefers "ineffective for *most*," but only because it is the most negative true framing, implying a lower value of θ than alternatives such as "ineffective for *some*." By contrast, a persuade-up speaker prefers "effective for *some*," highlighting the improved cases while obscuring the number of patients who did not improve.

Credulous and Vigilant Inferences by Listeners. How much does strategic underinformativity mislead listeners? A credulous listener who assumes informativeness is systematically biased. When a persuade-up speaker describes 2/5 patients as "effective for *some*," the credulous listener infers $\mathbb{E}[\theta] \approx 0.50$—substantially higher than the true proportion of 0.40. Conversely, a persuade-down speaker using "ineffective for *most*" leads the same listener to infer $\mathbb{E}[\theta] \approx 0.30$.

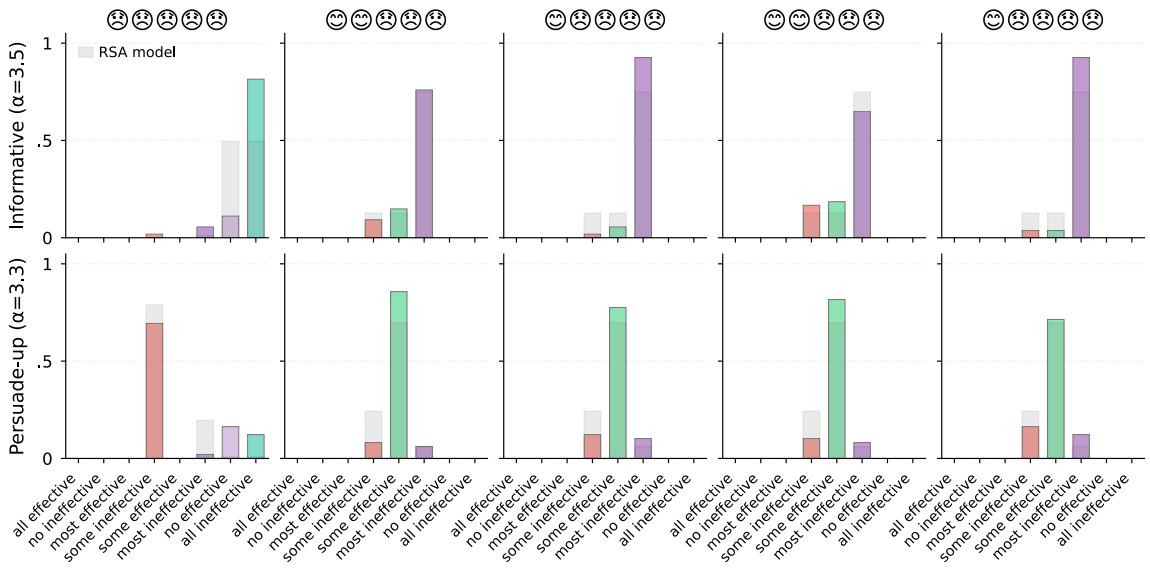

Figure 2: Speaker's Empirical Utterance Distribution and Model Prediction.

The gap between these inferences—despite describing identical observations—quantifies the "persuasion space" available to strategic speakers. A vigilant listener partially closes this gap. By maintaining uncertainty over ψ and reasoning about why the speaker chose that particular utterance, the vigilant listener's posterior is pulled back toward the truth. The intuition: if the speaker were truly informative, they might have been more precise; underinformativity itself is evidence of bias. Vigilance does not eliminate bias entirely—the listener cannot perfectly recover the speaker's goal—but it substantially reduces susceptibility to manipulation.

Learning Dynamic in Repeated Interactions. We ran $s = 5000$ simulated interactions over 60 rounds, with $n = 5$ patients per round and $\alpha = 3.0$. Figure 1B shows results for a true effectiveness rate of $\theta = 0.2$. When the speaker is informative, both credulous and vigilant listeners track the true value of θ. However, when the speaker is persuasive, the credulous listener shows a robust persuasion effect, whereas the vigilant listener remains comparatively resistant to persuasion and recovers estimates closer to the true θ.

Experiment 1: Speaker Production

Our simulations show that persuasive speakers *should* exploit underinformativity strategically. Do people actually produce these patterns? We tested whether human speakers shift their language use across informative and persuasive contexts.

Participants. We recruited 110 participants from Prolific (US). After excluding participants who failed attention checks or did not complete all rounds, 109 participants remained. Participants provided informed consent and were paid a base rate plus a bonus. Those who failed two attention checks were excluded and received prorated payment.

Design and Materials. Using a within-subjects design, participants described clinical trial outcomes under three communication goals presented in counterbalanced order: informative (describe accurately), persuade-up (convince the listener the treatment is effective), and persuade-down (convince the listener the treatment is ineffective). Stimuli depicted outcomes from a trial with $n = 5$ patients, each receiving one treatment session. Outcomes were displayed as a row of emoji faces whether the treatment was effective or ineffective for each patient. For each goal condition, participants completed 10 trials sampling from the space of possible observations (0–5 patients effective). On each trial, participants selected one utterance from the set of literally true descriptions. Utterances followed the template: "The treatment was [effective/ineffective] for [all/most/some/no] patients." Only true utterances appeared as response options.

Procedure. After consent, participants received instructions explaining the clinical trial scenario, quantifier semantics, and truth conditions. Comprehension was assessed through three modules testing: (1) definitions of *some* and *most*, (2) true/false judgments for specific utterance–observation pairs, and (3) identifying which observations satisfy a given utterance. Participants then completed the speaker task. For each goal condition, they were told that they were matched with another participant serving as the listener; in reality, listeners were simulated. One attention check was embedded at a random position within each block. After completing all three conditions, participants were debriefed.

Results. The model predicts that utterance distributions would diverge across communication goals. Informative speakers should prefer the most specific true description, regardless of whether the evidence favors or disfavors the treatment. By contrast, persuasive speakers should exploit underinformativity when the evidence conflicts with their goal. When observations indicate low effectiveness, persuade-up speakers should prefer underinformative true descriptions that highlight improvement or soften failure, such as "effective for *some*." Conversely, when observations indicate high

Model	Informative	Persuade-up	Persuade-down
Literal	2489.33	2526.73	2552.62
Informative	**1357.57**	2534.05	2560.03
Persuade-up	2496.41	**1934.07**	2560.06
Persuade-down	2496.56	2534.16	**1884.74**

Table 1: Model comparison across conditions using BIC. The best-fitting model in each condition is shown in bold.

effectiveness, persuade-down speakers should prefer underinformative descriptions that downplay improvement.

Human speakers exhibited patterns consistent with these predictions: when the observed outcomes conflicted with their persuasive goals, they shifted toward underinformative utterances rather than the maximally specific descriptions used in the informative condition (Fig. 2). For example, when no patients improved, informative speakers predominantly selected the most informative utterance, "all unsuccessful" (81.5%), whereas persuade-up speakers shifted toward the underinformative utterance "some unsuccessful" (69.4%).

Model comparison confirmed that goal-matched models provided the best fit in each condition. We fitted four models to human utterance choices: a literal speaker, an informative pragmatic speaker, a persuade-up pragmatic speaker, and a persuade-down pragmatic speaker. For each pragmatic model, the rationality parameter α was optimized to maximize log-likelihood. Table 1 presents model comparisons using the Bayesian Information Criterion (BIC). Across all conditions, the matched model provided the best fit. The informative speaker model achieved the lowest BIC in the informative condition, outperforming both literal and persuasive alternatives. The persuade-up model provided the best fit in the persuade-up condition, and the persuade-down model provided the best fit in the persuade-down condition.

Although the RSA models captured the overall pattern of goal-dependent language use, further inspection revealed systematic deviations attributable to processing costs and framing biases not included in the cost-free model. These deviations were most evident in equivalent utterance pairs: "all successful" is equivalent to "no unsuccessful" when all patients improve, and "no successful" is equivalent to "all unsuccessful" when no patients improve. Under the current RSA model, these pairs have equal utility, yet human speakers showed strong preferences between them. We decomposed these preferences (Figure 3) into two components: *baseline production preferences*, measured in the informative condition, and *goal-driven framing effects*, measured as deviations from this baseline in the persuasive conditions.

In the informative condition, where speakers had no strategic reason to prefer one framing over the other, we observed a strong preference for the "all" quantifier over "no." When describing uniformly effective outcomes, 100% of informative speakers chose "effective for *all*" over the equivalent "ineffective for *no*." When describing uniformly ineffective outcomes, 88% chose "ineffective for *all*" over "effective for *no*." This asymmetry is consistent with a production preference for universal affirmative descriptions over logically equivalent negative-quantifier descriptions.

Persuasive speakers deviated from this baseline in goal-congruent directions. When describing uniformly effective outcomes, persuade-down speakers showed an 86.1 percentage-point shift away from positive-predicate framing, decreasing from 100% "effective for *all*" in the informative condition to 13.9% in the persuade-down condition. This reversal suggests that speakers were willing to sacrifice baseline production preferences for persuasive impact. When describing uniformly ineffective outcomes, persuade-up speakers showed a 36.5 percentage-point shift toward positive-predicate framing, increasing from 12.0% to 48.5% choices of "effective for *no*." Both effects were significant (Fisher's exact tests, $p < .001$ for both comparisons), suggesting that speakers strategically adjust utterance choices beyond what the cost-free RSA model alone predicts.

Experiment 2: Listener Inference

Exp. 1 established that speakers strategically shift their language use based on communicative goals. Exp. 2 tested the complementary question: Do listeners reason pragmatically about speaker goals when interpreting utterances? Specifically, we asked whether listeners discount potentially biased language when they have reason to suspect that the speaker may be persuasive rather than purely informative.

Participants. Recruitment, exclusion, and payment were the same as in Exp. 1, 726 of 750 participants remained.

Design and Materials. We used a 3×3 between-subjects design crossing *listener condition* with *speaker condition*. The listener condition manipulated what participants were told about the speaker. In the *vigilant* condition, participants were informed that speakers could be one of three types: a Scientist, who provides informative descriptions; a Promoter, who favors the treatment; or a Skeptic, who disfavors it. Participants were told that they would not know which type they had been paired with. In the *credulous* condition, participants were told that speakers had been instructed to be "as informative and accurate as possible." In the *naturalistic* condition, participants received no information about the speaker.

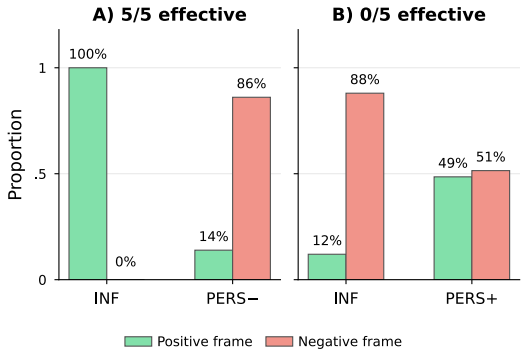

Figure 3: Baseline production preferences and goal-driven framing effects.

	Uniform Prior [.33, .33, .33]		Truth-default Prior [.25, .50, .25]	
Sequence	α	JS	α	JS
Informative				
Sequence 0	90.11	.148	14.18	.215
Sequence 1	1.10	.147	2.20	.153
Persuade-up				
Sequence 0	0.10	.258	0.10	.162
Sequence 1	0.10	.262	0.10	.205
Sequence 2	0.10	.203	0.29	.129
Persuade-down				
Sequence 0	0.10	.225	0.10	.214
Sequence 1	0.10	.232	0.12	.163
Sequence 2	0.10	.173	0.22	.124
Mean	—	.206	—	.171

Table 2: Model fit to speaker-type judgments in the vigilant condition with α fitted; lower JS distance indicate better fit.

The speaker condition determined which utterances participants actually received across five rounds. Utterance sequences were constructed to correspond to each speaker type. Informative sequences contained specific utterances, such as "effective for *all*" or "ineffective for *no*." Persuade-up sequences contained underinformative positive-leaning utterances, such as "effective for *some*." Persuade-down sequences contained underinformative negative-leaning utterances, such as "ineffective for *some*." Each speaker condition included two or three distinct sequences to ensure that results generalized across specific utterance patterns.

Procedure. Participants first learned about the clinical trial scenario and completed comprehension checks on quantifier semantics, following the same procedure as in Exp. 1. They then received listener-condition-specific instructions explaining that they would receive descriptions from a "speaker" partner who could see the trial results, but that they themselves would not see the outcomes directly. Instructions varied by listener condition as described above, but all participants were reminded that speakers could only produce literally true utterances. Participants were incentivized to estimate the treatment's effectiveness as accurately as possible.

To maintain the cover story of real-time interaction, participants saw a simulated pairing screen before the main task. On each of five rounds, participants viewed an utterance, such as "The treatment was **effective** for **some** patients," and estimated treatment effectiveness using a slider from 0% to 100%. Participants in the vigilant condition additionally indicated which speaker type they believed they had been paired with. After the main task, participants completed an attention check and final measures.

Results. We compared human responses to predictions from several listener models. The *literal listener* updates beliefs based solely on semantic content, without reasoning about speaker goals. The *credulous listener* assumes a cooperative speaker and reasons pragmatically about which observations would make the utterance likely. The *vigilant listener* maintains uncertainty over speaker types and jointly infers both treatment effectiveness and speaker goals.

We first fit the three listener models (literal, credulous, and vigilant) to participants' effectiveness estimates, optimizing the α parameter separately for each speaker condition × listener condition cell. Across all nine cells, the literal listener provided the best fit for informative speaker sequences (mean LL in $[-10.31, -7.91]$), whereas the credulous model provided the best fit for persuasive speaker sequences (mean LL in $[-14.40, -13.57]$). However, log-likelihood differences between competing models were uniformly small, typically less than 1, indicating that effectiveness estimates alone did not sharply distinguish the listener models. Critically, participants who were explicitly warned about potentially biased speakers showed the same general response pattern as those told the speaker was helpful or those given no information about speaker goals. This suggests that explicit knowledge about possible persuasive motives did not substantially alter how listeners interpreted utterances, at least as measured by point estimates of treatment effectiveness.

Participants in the vigilant condition also judged which speaker type (skeptic, scientist, or promoter) they believed they were paired with. We fit the vigilant listener model to these categorical judgments, first assuming a uniform prior over speaker types, $[.33, .33, .33]$. The first column of Table 2 presents the best-fitting α and Jensen-Shannon distance for each utterance sequence under this uniform prior.

Two findings emerged from this analysis. First, the model fit speaker-type judgments better for informative sequences than for persuasive sequences. This suggests that participants could identify informative speakers more readily than persuasive speakers, even though their effectiveness estimates were not sufficiently diagnostic to distinguish the listener models. Second, for all persuasive sequences, both persuade-up and persuade-down, the best-fitting α collapsed to near-floor values ($\alpha = 0.10$). Thus, even after hearing utterances such as "effective for *some*" or "ineffective for *some*" on four out of five rounds, participants showed little tendency to infer that the speaker was a Promoter or Skeptic, respectively. This suggests that listeners did not treat repeated underinformative utterances as diagnostic of persuasive intent. Instead, participants may have interpreted *some* as reflecting caution, uncertainty, or social appropriateness rather than strategic omission. This interpretation is consistent with open-ended responses such as, "I think they are a scientist because they use cautious words such as 'some'."

Lastly, when looking at the pattern in the posterior for persuasive speakers across persuasive conditions, we find that participants showed inflated posterior credence for informative speakers, suggesting a potential "truth-default bias" (Levine, 2022). To test the possibility of this bias, we fitted a model with the prior assigned to the informative speaker as a parameter, and recovered a prior of $[.25, .50, .25]$ reflecting a "truth default" assumption that speakers are more likely to be informative than biased. The second column of table 2 presents the best-fitting α and Jensen-Shannon distance for

each utterance sequence under the truth-default priors. However, even under the best-fitting prior, the optimal α values for speaker-type inference remained near floor (0.10–0.29). This indicates that listeners were not systematically using utterance patterns—such as the choice of "some" over more precise alternatives—to infer whether the speaker's goals.

Overall, contrary to the model's prediction, listeners did not treat underinformativity such as *some* as diagnostic of persuasive goals. This remained true even when participants were explicitly informed that speakers might be biased.

Discussion

Selective truth-telling is a pervasive tactic in strategic communication. Our results reveal an asymmetry in how this phenomenon manifests on the production and comprehension sides of communication. On the production side, when multiple true utterances are available, persuasive speakers systematically select utterances that bias listener inference toward their goals—not by lying, but by strategically deploying underinformativity. Exp. 1 confirmed this prediction: speakers shifted toward underinformative quantifiers such as "some" when persuading compared to informing. This pattern emerges naturally from an RSA model in which speakers maximize goal-dependent utility over the true utterances.

However, we also observed additional biases that go beyond the rational speaker account. First, speakers showed framing effects: they were more likely to mention effective outcomes when persuading up and ineffective outcomes when persuading down (Tversky & Kahneman, 1981), even when these utterances had the same utility under the model. This suggests that speakers may have direct preferences over which aspects of reality to highlight, beyond what is captured by reasoning about listener inference. Second, we found evidence consistent with cognitive processing costs: speakers preferred simpler or more accessible utterances (Carpenter & Just, 1975; Clark & Chase, 1972). Together, these findings suggest that RSA captures the core pattern of strategic underinformativity, while human speakers layer additional framing and processing biases on top of this rational baseline.

On the comprehension side, the RSA framework predicts that vigilant listeners should be able to detect persuasive intent by reasoning about *why* a speaker chose a particular utterance. A speaker who repeatedly uses underinformative descriptions, such as "effective for *some*" across multiple trials, should reveal something about their goals—or so the model predicts. Yet Exp. 2 found only limited evidence for such vigilance. Listeners showed modest sensitivity to quantifiers such as "most" and "all" when inferring speaker type, but virtually no sensitivity to "some," the quantifier most useful for selective truth-telling. Critically, even participants who were explicitly warned about potentially biased speakers and asked to judge speaker type did not reliably use underinformativity or one-sided framing as cues to persuasive intent.

This asymmetry suggests that the symmetric recursive reasoning assumed may break down in strategic contexts. Speakers appear to exploit pragmatic inference by selecting utterances that listeners are likely to interpret favorably. But listeners do not appear to reciprocate by asking why speakers chose those utterances rather than more informative alternatives. This creates a potential vulnerability: if speakers can strategically frame information while listeners interpret utterances at face value, selective truth-telling may be more effective than a symmetric model would predict.

Why might this asymmetry arise? One possibility is that epistemic vigilance is cognitively costly (Sperber et al., 2010). Maintaining uncertainty over speaker types and updating beliefs based on utterance patterns requires tracking information that may rarely be useful in cooperative communication. If most everyday communication is cooperative, listeners may interpret utterances charitably by default—a "truth default" that serves them well on average but leaves them vulnerable to strategic speakers (Levine, 2022).

Another possibility is that listeners have a different pragmatic interpretation of underinformativity than the model assumes. Open-ended responses suggest that some participants interpreted "some" not as a sign of persuasion, but as a marker of epistemic caution. In the context of medical treatments and scientific communication, hedged language may signal that a speaker is being appropriately careful rather than strategically underinformative. A scientist who says "some patients improved" might be seen as responsibly avoiding overconfidence, not as hiding unfavorable information. This alternative pragmatic construal—where underinformativity signals epistemic humility rather than bias—would make "some" a poor diagnostic cue for detecting persuasive speakers, even for listeners who are actively reasoning about speaker goals.

A related possibility concerns the constrained nature of the experimental utterance space. Listeners in Exp. 2 only experienced the comprehension side of communication; they never served as speakers and thus had no direct experience with the choice set available to speakers. Without this "speaker's-eye view," listeners may not have appreciated how informative the choice of "some" was relative to alternatives such as "most" or "all." In natural communication, listeners may have implicit knowledge of what speakers *could* have said, built up through experience as both speakers and listeners. Our paradigm, by separating these roles across experiments, may have deprived listeners of the experiential basis for vigilant inference. This suggests that epistemic vigilance may emerge more readily in contexts where individuals have experience on both sides of the communicative exchange.

Finally, the cues to persuasive intent in our experiment may have been too subtle. Underinformative quantifiers such as "some" are common in ordinary speech, and repeated use may not strongly signal bias without additional contextual support. Listeners may become more vigilant when other cues to speaker motivation are present, such as explicit persuasive framing, prior experience with a particular speaker, or richer information about the process by which speakers observe and report evidence (cf. Cummins & Franke, 2021).

Acknowledgments

CMW is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (C^4: 101164709), the Hessian research funding programme LOEWE/4b//519/05/01.002(0022)/119, the Deutsche Forschungsgemeinschaft (German Research Foundation, DFG) under Germany's Excellence Strategy (EXC 3066/1 "The Adaptive Mind", Project No. 533717223), and the Excellence Cluster "Reasonable AI" by the Deutsche Forschungsgemeinschaft (German Research Foundation, DFG) under Germany's Excellence Strategy – EXC-3057.

References

Carpenter, P. A., & Just, M. A. (1975). Sentence comprehension: A psycholinguistic processing model of verification. *Psychological Review*, *82*(1), 45–73.

Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology*, *3*(3), 472–517.

Corriveau, K. H., Meints, K., & Harris, P. L. (2009). Early tracking of informant accuracy and inaccuracy. *British Journal of Developmental Psychology*, *27*(2), 331–342.

Crawford, V. P., & Sobel, J. (1982). Strategic information transmission. *Econometrica: Journal of the Econometric Society*, 1431–1451.

Cummins, C., & Franke, M. (2021). Rational interpretation of numerical quantity in argumentative contexts. *Frontiers in Communication*, *6*, 89.

De Leeuw, J. R., Gilbert, R. A., & Luchterhandt, B. (2023). Jspsych: Enabling an open-source collaborative ecosystem of behavioral experiments. *Journal of Open Source Software*, *8*(85), 5351.

Degen, J. (2023). The rational speech act framework. *Annual Review of Linguistics*, *9*(1), 519–540.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829.

Kamenica, E., & Gentzkow, M. (2011). Bayesian persuasion. *American Economic Review*, *101*(6), 2590–2615.

Levine, T. R. (2022). Truth-default theory and the psychology of lying and deception detection. *Current Opinion in Psychology*, *47*, 101380.

Oey, L. A., Schachner, A., & Vul, E. (2023). Designing and detecting lies by reasoning about other agents. *Journal of Experimental Psychology: General*, *152*(2), 346.

Oey, L. A., & Vul, E. (2024). Accurate approximations about the truth from literally false messages. *Computational Brain & Behavior*, *7*(1), 23–36.

Rogers, T., Zeckhauser, R., Gino, F., Norton, M. I., & Schweitzer, M. E. (2017). Artful paltering: The risks and rewards of using truthful statements to mislead others. *Journal of Personality and Social Psychology*, *112*(3), 456.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, *25*(4), 359–393.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458.