Computational Brain & Behavior
https://doi.org/10.1007/s42113-023-00187-0

ORIGINAL PAPER

Accurate Approximations About the Truth from Literally False Messages

Lauren A. Oey[1,2] · **Edward Vul**[1]

Accepted: 18 October 2023

Abstract

Communication can be weaponized to manipulate others' beliefs, most glaringly via explicit lies. We investigate one defense mechanism: people can infer the truth from false messages when they expect that (1) speakers have adversarial motives that direct their lies and (2) bigger lies are costlier. We show in a lab experiment that people can correct for bias in lies when these conditions are satisfied, but with decreased precision. When people adjust what information they glean from expected dishonesty, how might this perturb dyadic, and moreover collective, communication channels? Through probabilistic simulations, we find that deceptive communication systems converge to equilibrium states, in which listeners extract accurate (but less precise) estimates of the truth. Furthermore, when listeners correct for messages assuming that they are distorted, even cooperative speakers (who want listeners to have the correct interpretations) should lie. Liars do not get their way, but they make communication noisier for listeners and other speakers.

Keywords Deception · Communication · Lie detection · Social cognition · Probabilistic models

Communication generally aims to faithfully transmit information from a speaker's mind to a listener's mind. Typically, the listener expects the speaker to be honest, and the speaker expects to be interpreted as honest (Grice, 1975). The aligned goals allow for people to effortlessly converge on a shared communication system (Clark, 1996; Tomasello et al., 2005). However, there are many forms of communication that feature misaligned goals and call upon listeners to be vigilant, such as in deception (Sperber et al., 2010). If cooperative communication aims for the listener to infer an accurate depiction of what the speaker thinks, one aim of deceptive communication is to induce in the listener a distorted depiction. Broadly, this highlights information transmission as a key incentive for both cooperative and deceptive communication.

Consider a speaker who strategically lies – in doing so, they want to distort the listener's belief, rather than simply trying to remain undetected. Similarly, the listener wants to extract an accurate representation of the truth from the distorted message, rather than simply sleuthing out whether a message is a lie or not. Thinking of deceptive communication as information transmission couches listeners as lie *interpreters*, asking people what meaning they extract out of a message. This perspective deviates from the traditional focus of listeners as lie *detectors*, who categorize messages as true or false (e.g. Bond and DePaulo, 2006, 2008; ten Brinke et al., 2016; Levine et al., 1999; Leach et al., 2004; Oey et al., 2023). Here, we examine inference when deceptive intent is already suspected, so listeners are not burdened with worrying about whether a message is a lie – rather, what is the truth, given that this message is likely a lie. In doing so, we introduce a formulation of the goals of deception that emphasize a *social intention* to manipulate others' beliefs.

The premise of transmitting distorted messages occurs across numerous human communication systems. One such communication system is letters of recommendation, in which letter writers seek to promote their candidate, so they are highly motivated to inflate their candidate's apparent qualifications. However, they face constraints — for example, baldfaced over-embellishments may hurt letter writers'

Revision of COBB-D-23-00026 as invited by the Editor in Chief, Scott Brown, PhD.

✉ Lauren A. Oey
loey@princeton.edu

[1] University of California, San Diego, Department of Psychology, San Diego, California, USA

[2] Princeton University, Department of Psychology, Princeton, NJ, USA

Published online: 20 November 2023

reputation. Meanwhile, letter readers want to accurately assess the candidate's qualifications. The asymmetric goals between letter writers and readers, and asymmetric knowledge that writers know more about the ground truth than readers, promotes the passing of distorted messages. Qualified candidates are not simply "good," they are "the complete package."

Given that letters of recommendation are fraught with embellished language, the communication system at first glance seems prone to erroneous information transmission, like how learners who are exposed to biased samples tend to draw biased inferences (Hogarth et al., 2015; Feiler et al., 2013). Yet, our continued use of letters of recommendation superficially suggests that letters are in large scale effective at communicating information. And for individual readers to value using letters, they must expect to extract meaningful information. Just as learners who are aware of the sampling constraints correct their generalizations (Hayes et al., 2019), perhaps readers systematically correct for biased language in letters. This raises a puzzle at the level of dyadic communication: how do people interpret distorted messages?

On one hand, people may rely on domain-specific, established conventions that give rise to the meaning of messages (Lewis, 1969). A distorted message serves as one arbitrary solution to the coordination problem on meaning. Both speakers and listeners align on mapping the message ("the complete package") onto the interpretation ("good"). However, this serves as a dissatisfying explanation to how distorted communication systems arise because a seemingly more salient solution would be to transmit truthful messages (Schelling, 1960; Lewis, 1969), as opposed to distorted ones.

Alternatively, people may interpret distorted messages guided by domain-general mechanisms, namely rational theory-of-mind reasoning (Oey et al., 2023). The core mechanism driving people's lie interpretation is an assumption about speakers' goals and an intuitive understanding of how goals drive speakers' behavior. Broadly, rational theory-of-mind frameworks are grounded on the assumption that both speakers and listeners generate decisions that maximize their rewards, and they intuit that other agents do the same (Jara-Ettinger et al., 2016; Baker et al., 2017). Recent implementations have proven useful for explaining numerous speaker-phenomena in deception, such as how suspicion influences people to preferentially mislead or be uninformative (Franke et al., 2020; Ransom et al., 2019), or how plausibility drives the extremeness of lies (Oey et al., 2023). In interpreting communicative messages too, listeners draw meaning about what was said guided by their assumptions of speakers' goals. When listeners' assumptions about speakers are misplaced, listeners are vulnerable to being deceived. For example, in pedagogy, learners assume that knowledgeable teachers demonstrate a concept by being fully informative. If teachers omit information, learners can be misled into thinking they know all that there is to know about the concept, when this is in fact a false conclusion (Bonawitz et al., 2011). When made aware of the teacher's tendency to omit information, even children can adjust what meaning they draw from messages, such as "there is more to learn about this concept than what I have been told" (Gweon et al., 2014). Therefore, listeners seem to be equipped with cognitive tools to be vigilant to what information they receive.

Our work proposes that vigilant listeners can make richer inferences than those seen in the previous literature. Ushered by rational assumptions, a listener, who observes a speaker's distorted message, may reverse engineer *what the speaker thought was the truth*. A key prediction of this framework is that people can robustly tune how they interpret the truth to their knowledge about the speaker's goals. In the teacher example, not only might listeners infer that there is more to learn, rather more specifically, sufficient knowledge about the speaker will invite listeners to infer that there is *one* more thing to learn, such as the teacher withholding information to test generalization about a concept. In contrast, learners could also suspect that there are *many* more things to learn because the teacher is withholding a substantial amount of information, as to encourage learners to explore more independently. Crucially, rational theory-of-mind accounts do not automatically bestow omniscience – the accuracy of listeners' inferences about the truth depends on the veracity of their mental model of the speaker, such as how they conceptualize speakers' incentives and costs to lie.

Related work has looked at perfectly rational agents in game theoretic models to examine what messages are made when listeners suspect deceptive speakers. Classic conclusions from this literature propose that at equilibrium speakers ought to provide an information-less message and listeners ought to assume that no information is contained within the message (Crawford & Sobel, 1982). However, empirical tests of this proposal have found that speakers provide surprisingly more information and listeners select actions by considering speakers' messages more than they should in theory (Cai & Wang, 2006). By introducing a simple principle – that speakers' lies are related to their incentives and costs – Kartik (2009) elegantly shows that this property of behaviors can emerge as a stable equilibrium state in these models. Underlying this model is an assumption that listeners have perfect access to speakers' incentives and rationally adjust their actions accordingly. The result is a number of interesting predictions about fully rational behavior, which we revisit in this paper. First, the form that the equilibrium state takes is systematically pushed around as a function of the conflict between speakers' and listeners' incentives and as a function of speakers' costs. Second, speakers are incidentally disincentived from telling the truth or generating smaller lies because "it would lead to adverse inferences from the receiver, who expects the equilibrium degree of language

inflation and thus rationally deflates." We see our own work as a complementary application of Kartik (2009). Importantly, we empirically test an assumption of Kartik (2009), that listeners can and do integrate what they believe about both speakers' incentives and costs to generate reasonable inferences about the truth. Additionally, we expand on this previous work by considering the properties of a communication channel operating under such incentives, including (a) what inferences would listeners make about reality, (b) how would speakers motivated to be honest communicate, and (c) what is the overall bias and noise of the resulting channel.

For distorted communication systems to arise, we presume that there are two necessary preconditions that constrain how goals influence messages. (1) Speakers want to induce the listener into believing something about the world that is not only literally false, but is also favorable to the speaker. For listeners to tune their inferences, they need to be aware of the speaker's broad goals. (2) Speakers must face costs to constructing lies that deviate from the truth. For instance, process models, such as those that propose a direct relationship between speakers initially thinking about the truth and secondarily manipulating the truth to produce a lie (e.g. Walczyk et al., 2014; Debey et al., 2014) broadly predict that lies that more distantly deviate from reality require more cognitive effort to construct. These costs might be due to increasing risk of detection (Oey et al., 2023), loss of plausible deniability (Pinker et al., 2008), higher cognitive load (van't Veer et al., 2014), managing reputation (Abeler et al., 2019), other-regarding preferences (Gneezy, 2005; Maggian & Villeval, 2016), or moral signaling (Gneezy et al., 2018). Regardless of the exact forces at play within a given individual or context, speakers' costs broadly drive speakers to produce smaller lies that are closer to reality (Lundquist et al., 2009; Shalvi et al., 2011; Gneezy et al., 2018; Gerlach et al., 2019).

Deploying a novel behavioral task, we tested if people successfully infer the truth when the two preconditions are met. A rational theory-of-mind framework not only predicts that people should succeed with sufficient knowledge about the speaker, but it uniquely predicts that people tune their truth inferences to different beliefs about speakers' goals and costs. Participants played a game in which a sender draws red and blue marbles from a jar and sends a manipulated representation of their marbles to a judge by clicking marbles on the interface (a physical cost to produce larger lies). Seeing the manipulated representation, the judge guesses how many red marbles were truly drawn. We found that people robustly interpreted distorted messages in a way that was tuned to speakers' directional goal and the magnitudinal cost.

Applying probabilistic models, we tested the downstream consequences of listeners' lie interpretations. The rational theory-of-mind framework sets up a first principles approach to simulate how agents' goals influence distorted communication systems. One possibility is that speakers (and listeners) plan their behaviors attempting to out-strategize the other. The result is an arms race between speakers ratcheting up the extremeness of their distorted messages and listeners ratcheting up their corrections, so that "letters of recommendation" become increasingly detached from reality. Contrary to this intuition, by applying probabilistic simulations, we showed that messages and interpretations converge to an equilibrium state when listeners suspect speakers' goals. Rewards and costs of speakers influence the accuracy and precision of listeners' truth inferences: as the cost to lying decreases relative to the reward for deceiving a listener, the transmission of the truth remains unbiased (though more imprecise). Furthermore, when listeners generalize their suspicion to others, the consequence is that other speakers are indirectly affected. Cooperative speakers, wanting to guide listeners to accurate beliefs, are pressured to say *dishonest* messages when they expect vigilant listeners.

Overall, our study informs our scientific understanding of how distorted messages, but nonetheless faithful transmission, can perpetuate in communication systems. Our probabilistic, goal-based paradigm reveals that distorted communication systems are not simply odd pockets of anomalies, rather they commonly occur and produce systematic behaviors. Underlying these communication systems are people's robust ability to engage in rational theory-of-mind reasoning, which powers people to extract clairvoyant insight about the truth from falsehoods.

Human Experiment: Testing Goals and Costs as Preconditions to Infer the Truth

As an initial proof of concept, we tested if people not only infer the truth from lies, but they robustly tune their inferences to the speaker's goals and costs, in an experimental setting where the preconditions apply. Namely, that they are aware of the speakers' motive to directionally bias their lie, and that they face costs for producing more extreme lies.

Methods

Participants

Participants were recruited from the undergraduate population at the University of California, San Diego to participate in an online game for course credit. Data was collected from 254 participants. Of these, 44 participants were excluded for failing to answer at least 75% of the attention check questions within a ± 5 error, and six participants produced multiple responses that were out-of-bounds. Participants who produced a single out-of-bounds trial had that trial excluded from

analysis, but their remaining trials were included. In total, 204 subjects were included in our final data set. Informed consent was obtained from all participants, and the study was approved by the university's Institutional Review Board.

Procedure

Participants played an adversarial communication game, alternating between the roles of sender and judge. Senders saw a display of 100 red and blue "marbles" arranged in a 2D jittered grid, reflecting the ground *truth* sampling of red and blue marbles from a virtual jar (Fig. 1). The sender could alter how many red (and blue) marbles were in the display by manually clicking individual marbles to swap their color, before sending to the judge the *message*, the altered snapshot of the marbles. The judge, in turn, sees the shaken display of marbles and the number of red marbles in it (i.e. the message), and then has to *estimate* the original, ground-truth number of red marbles.

The players' goal was to win against the other player by the largest possible point differential. Judges lost points corresponding to the absolute (L1) error of their estimate, so in Fig. 1, a guess of 50 when the truth was 48 resulted in −2 points. Meanwhile, senders gained points for the judge's error in the direction of the sender's goal, so a sender who wanted the judge to *overestimate* got +2 points. If the judge guessed in the opposite direction (e.g. underestimated instead), the sender got 0 points, but the judge still got −2 points for their absolute error.

The critical between-subject manipulations were the *goals* and *costs* of deception for the sender. Senders were assigned the goal to make the judge either *Overestimate* or *Underestimate* the number of red marbles, while judges aimed to accurately guess the truth. The number of clicks required to

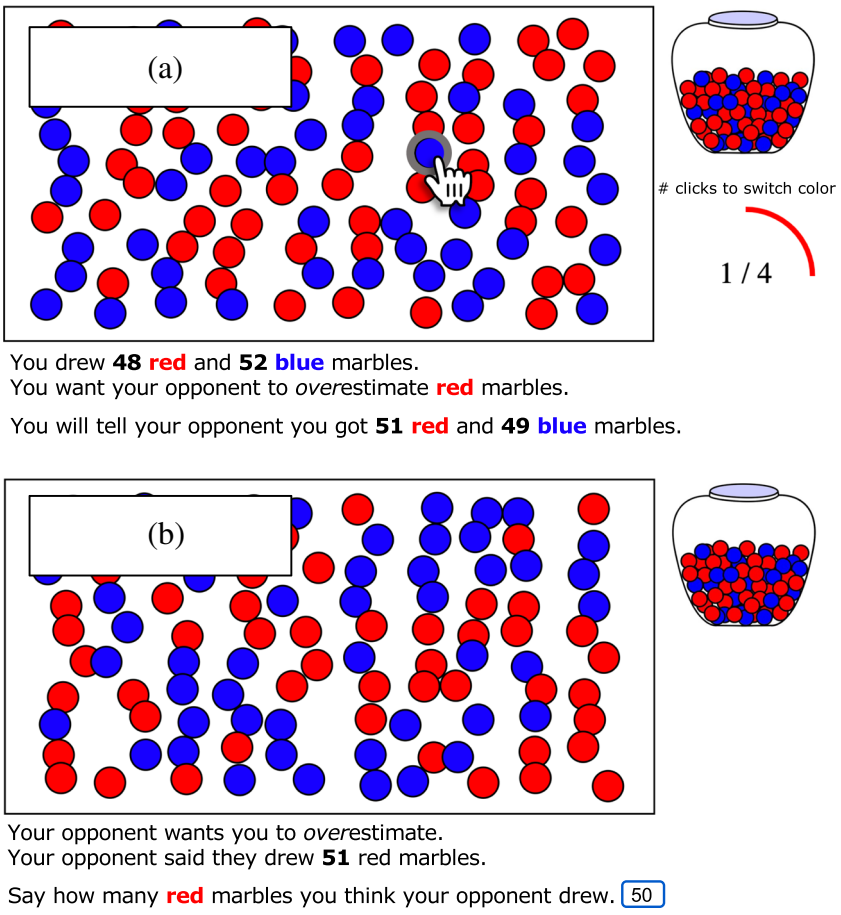

Fig. 1 Game design. (**a**) The sender sampled marbles, and could manipulate what they showed their opponent about how many red marbles they drew by clicking marbles in the display to flip their color. The sender is told in text how many of each color marble they originally drew (e.g. "You drew 48 red...") and how many they would currently report based on their clicks (e.g. "You will tell your opponent you got 51 red..."), and a progress bar shows how many more clicks are needed to switch the next marble. (**b**) The judge tried to estimate how many marbles the sender truly drew from what the sender reported. In this example, the sender wants the judge to *Overestimate*, and producing larger lies follows a `Quadratic` cost function (requires additional click for each additional flip). Here, the *truth* was 48 red marbles, but the *message* was a lie of 51. The judge *estimated* the truth to be 50

change the color of a marble served as a physical cost for the sender to generate more extreme lies. Specifically, participants either needed to click each marble just once to switch their color (lower `Linear-cost`), or they needed to click each marble an additional time to switch color resulting in the number of required clicks to grow quadratically with n: $1 + 2 + 3 + ... + n = \frac{n(n+1)}{2}$ (higher `Quadratic-cost`). Thus, participants in the `Quadratic-cost` condition needed to exert manual effort to produce more extreme lies. If the amount of effort senders committed to trials was consistent between the conditions, then we would expect that senders produce less bias in their lies when subjected to higher costs in the `Quadratic-cost` condition. Participants were randomly assigned to one of the 2×2 conditions.

Participants were explicitly instructed about the goal of the sender, and the cost to switch marbles (e.g. `Quadratic-cost`: "The more marbles you switch color, the more clicks you'll need to switch each marble.") During `Quadratic-cost` sender trials, a circular progress bar tracked the number of clicks already completed and the number of additional clicks to switch a given marble color. In the `Linear-cost` condition, the circular progress bar was not present. Participants were instructed that the original jar was composed of 50% red and 50% blue marbles. However the marbles were sampled from a beta-binomial distribution $X \sim BetaBinomial(100, 3, 3)$ (95% of samples fall between 14 and 86), which while still centered at 50, yields more variability than a standard binomial distribution (95% fall between 40 and 60). By increasing the variability of ground-truth, we expected that participants would rely more on their beliefs about cost functions, rather than base-rates, to judge the truth. To avoid counting errors, both the sender and the judge were also explicitly informed of the number of red and blue marbles in the display. Furthermore, to avoid concerns about the positional distribution of marbles serving as a cue to deception, the positions of marbles were shuffled before the senders' altered display was shown to the judge.

Participants played against a computer opponent, which allowed us to control for the opponent's behavior. Participants were not explicitly told if their opponent was a computer or a human. The computer opponent's response time, average lying, and inference behavior was held constant across cost function conditions to ensure that any potential variation in participants' judgments of ground-truth was caused by their beliefs about the sender's cost function and not by the computer sender's actual behavior. Specifically, the computer sender lied by taking the truth and adding in the direction of their goal some sampled amount, taken from a Poisson distribution with a mean of 5. As a judge, the computer sampled from the same distribution but subtracted from the participant's message.

Participants played for two practice trials: first as the sender, then as the judge. Then, participants played for 100 test trials, alternating between sender and judge roles every trial (which role was played first during the test trials was randomized). Throughout the task, participants additionally answered 12 attention check questions related to the trial (two in the practice trials, and ten randomly distributed in the test trials). To prevent participants from relying on learned information about their opponent's behavior, participants did not receive direct feedback about their opponent's decision or the trial's outcome. Instead, they received feedback about the players' cumulative points every five trials, which motivated participants to play the game while only revealing coarse information about their success.

Results

Validating Preconditions in Lying Behavior

We first validated that the condition manipulations worked, and senders chose lies that were driven by their assigned goals and were systematically constrained by the assigned cost function. Senders biased their lies in the same direction as their goal to induce the judge to over- or underestimate (Fig. 2). Using linear models with random-effects for subject and item (the true draw), we found that (as expected) senders whose goal was to overestimate inflated their message relative to the truth ($\hat{\beta} = 5.98, t(162) = 5.99, p < 0.0001$), and those whose goal was to underestimate deflated their message ($\hat{\beta} = -4.46, t(132) = -5.50, p < 0.0001$). Additionally, although the bias point estimate is systematically larger for senders with the goal to overestimate, there is not a significant difference. We also validated that the cost conditions systematically influenced how senders lied: `Linear-cost` senders introduced more bias into their message relative to the `Quadratic-cost` senders ($\hat{\beta} = 5.15, t(202) = 4.81, p < 0.0001$), aggregating over goals. These results showed that senders generated lies consistent with their assigned goal and cost function.

Do People Estimate the Truth by Considering their Beliefs About Others' Goals and Costs?

Judges who apply their beliefs about senders' goals should make bias corrections in the opposite direction of the senders' goal. If the sender wanted the judge to overestimate, then the judge should expect the sender to positively bias their message by adding more red marbles. A judge who expects positive bias in the message should correct for the bias in the negative direction by estimating that the true number of marbles drawn was fewer than what was reported in the message. As predicted, we found that participants in the *Overestimate* condition bias-corrected in the negative direction by guessing smaller numbers ($\hat{\beta} = -6.84, t(148) = -7.31, p < 0.0001$). Vice versa, participants in the *Underestimate*

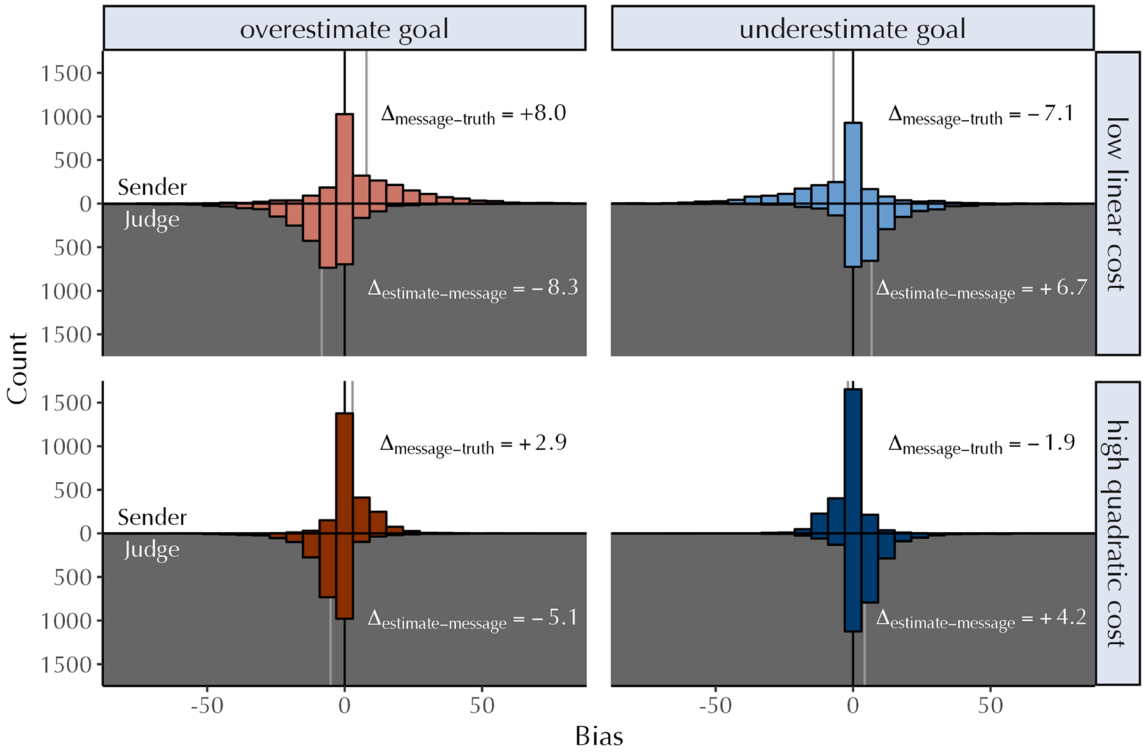

Fig. 2 Distribution of participant senders' biasing and judges' bias-correcting behavior across each condition (panel columns are goals and rows are cost functions). The direction and distance of the gray line (mean bias) relative to the black line at $bias = 0$ indicates if participants generally inflate (positive bias) or deflate (negative bias) their response and how large the difference is. The top half of each panel (in white) shows how much senders manipulate their message relative to the truth ($\Delta_{message-truth}$). Senders with *Overestimate-goals* (left panels) biased their messages in the positive direction from the truth, and vice versa, senders with *Underestimate-goals* (right panels) biased their messages in the negative direction. Senders with lower `Linear-costs` produced more bias (mean is farther from 0) in their message, than those with higher `Quadratic-costs`. The bottom half of each panel (in gray) shows how much judges adjust their truth estimate relative to the sender's message ($\Delta_{estimate-message}$). Judges bias corrected in the opposite direction of senders' bias – when judges expected senders to have *Overestimate-goals* and bias their message in the positive direction, judges tuned how they bias corrected their estimate in the negative direction. Judges also tuned how they bias corrected to the sender's expected costs – when judges expected senders to have lower `Linear-costs` to lie, they bias-corrected more (mean is farther from 0) in their estimate of the truth. The mean bias for each role and condition is shown in text on the plot

condition bias-corrected in the positive direction by guessing larger numbers ($\hat{\beta} = 5.33, t(159) = 7.20, p < 0.0001$). Once again the bias correction point estimate is systematically larger for receivers in the overestimate goal condition, but there is not a significant difference. These results show that the direction of people's truth inferences are informed by their beliefs about speakers' goals.

Judges that apply their beliefs about senders' cost functions should expect senders with lower cost functions to produce more extreme lies. Therefore, they should make larger magnitude bias corrections. Indeed, judges who believed the sender had a `Linear-cost` debiased their estimate more compared to the `Quadratic-cost` ($\hat{\beta} = -2.96$, $t(202) = -3.53$, $p < 0.001$), aggregating over goals. Figure 2 shows that the bias corrections' absolute distance to the intercept is larger for the `Linear-cost` (top panels), compared to the `Quadratic-cost` (bottom panels).

Lastly, when comparing human judge bias correcting to human sender biasing, we do not see any systematic over- or undercorrections. People are not overly nor insufficiently trusting relative to how people would lie in this task. Thus, people broadly seem to calibrate how they correct for bias to how they add bias into their lies. A more thorough investigation into individual differences between individuals' own sender and judge behavior is included in the Supplementary Materials.

We asked whether people can estimate the truth from the content of a lie. We tested the hypothesis that this feat can be achieved without clairvoyance so long as listeners know how speakers are (1) directionally motivated to lie, and (2) cost-constrained in the magnitude of their lies. Our behavioral experiment manipulated the goals and costs of speakers' deception, and showed that participants are sensitive to these factors when lying. Critically, people are also calibrated to

the senders' goals and costs when they try to estimate the truth from the content of the lie, suggesting that in settings where goals and costs are transparent, overt lies may not actually lead to systematic deception.

Probabilistic Simulations: Consequences for Communication Systems

Our behavioral study showed that listeners can estimate reality from deceptive messages by considering the speaker's motives and costs. The behavioral result suggests that in certain communication channels, senders and judges pass around dishonest messages, yet judges approximately infer the ground truth. This result opens a number of questions about how communication would work in such settings. Consider again recommendation letters. First, if a letter writer predicts readers take away a softened interpretation of writers' claims, then perhaps a writer ought to further amplify their claims about the candidate. If such escalation proceeds unchecked, recommendation letters may become completely decoupled from reality. What are the requirements to keep this process in check, and what properties do we expect of the resulting communication channel?

Second, while some letter writers may embellish their claims, other writers may want to accurately convey their beliefs about a candidate. When there is a mixture of speakers who have varying motivations, listeners may be best served by assuming the speaker is semi-deceptive, semi-cooperative and systematically curb their vigilance about reality accordingly. Under this assumption of listener behavior, dishonest messages will be interpreted nonliterally, and so too will honest messages. Thus, a cooperative speaker, who intends for the listener to extract an accurate interpretation, will fall short if they simply say an honest message. How would cooperative speakers behave in an environment with listeners that expect many deceptive speakers? In the next section, we examine these population-level dynamic using probabilistic modeling.

Model Setup

We consider two interacting agents: senders and judges. Senders observe some ground truth (k) and select a message to say to the judge (k_{say}); thus they are characterized by their utterance distribution $P_S(k_{say} \mid k)$. Judges observe the message from the sender (k_{say}) and produce an estimate of the truth (k_{est}), and so are characterized by their estimation distribution $P_J(k_{est} \mid k_{say})$. The conditional response distributions of senders and judges arise from a decision rule over their expected utilities, calculated from their utility functions. To maintain generality, these utility functions are both defined over $\{k, k_{say}, k_{est}\}$ tuples.

In the basic agent model we consider here, judges want to accurately estimate the truth, so their utility function can be characterized as an L2 loss function on the error of k_{est} relative to k without considering the specific message they received (k_{say}) at all:

$$U_J(k, k_{say}, k_{est}) = -(k_{est} - k)^2 \tag{1}$$

While judges have a simple, constant goal to be accurate, it is useful to consider senders with different goals. Broadly, pragmatic senders design a message about the world by considering what beliefs it will instill in the judge.

Deceptive senders (S_D) aspire to mislead the judge by causing them to mis-estimate the truth. This can be captured by utility that scales with the error of the judge's estimate (k_{est}). However, this deceptive sender does not wish to produce messages too far off from reality because of a cost function penalizing increasing falsity in their message. In the behavioral experiment we directly manipulate lying magnitude cost in terms of manual effort to make the manipulation experimentally tractable; in the real world, costs to lying are imposed by a broad set of cognitive and social norms and constraints, such as the cognitive effort to generate a large yet still plausible lie, to maintain plausible deniability about one's intent, or to deviate from one's own intrinsic aversion to being dishonest. Our models abstract all such cognitive and social costs into one mathematical term, as in Kartik (2009). These costs can be captured by an L2 loss on deviations between message and reality.

$$U_{S_D}(k, k_{say}, k_{est}) = (k_{est} - k) - m(k_{say} - k)^2 \tag{2}$$

Note that we have chosen a L2 (quadratic) loss instead of a linear loss function, which would have resulted in peculiar maximal lying or no lying at all behavior depending on the slope of the desire to induce a biased inference versus the slope of the cost. The parameter m represents a ratio of the deceptive sender's relative desire to induce a biased inference in the listener versus their cost to make messages more discordant from reality.

In contrast, pragmatic cooperative senders (S_C) have goals that align with judges, and thus also want judges to form accurate beliefs about the world, and only consider an L2 loss function on the judge's estimate error:

$$U_{S_C}(k, k_{say}, k_{est}) = -(k_{est} - k)^2 \tag{3}$$

This pragmatically-cooperative utility function is notably different from that of a literally honest sender, who would only seek to minimize the deviation of their message from reality ($-(k_{say} - k)^2$) regardless of how that message is understood by the judge. This distinction between considering how a message is interpreted, rather than its literal

meaning, is at the heart of modern models of cooperative communication (Frank & Goodman, 2012; Goodman & Frank, 2016), which argue that human language use can be understood in terms of such pragmatic motives. Later, we will see this distinction between pragmatic, and literal, honesty is important when cooperative speakers share a communication channel with liars. Other utility terms may be considered, but for our purposes, this is a minimal set to illustrate the dynamics that emerge in not-entirely-cooperative communication channels.

The decision rule for the sender and the judge are given as the softmax of their expected utilities, where α is the decision noise parameter. Defining these decision rule entails mutual recursion because the sender's utilities depend on the predicted response of the judge

$$P_S(k_{say} \mid k) \propto exp(\alpha \sum_{k_{est}} U_S(k_{est}, k_{say}, k) P_J(k_{est} \mid k_{say})) \tag{4}$$

and the judge's utilities depend on inverting the sender's message distribution

$$P_J(k_{est} \mid k_{say}) \propto exp(\alpha \sum_{k} U_J(k_{est}, k_{say}, k) P(k \mid k_{say})) \tag{5}$$

where the conditional probability of the truth is given by:

$$P(k \mid k_{say}) = \frac{P_S(k_{say} \mid k) P(k)}{\sum_k P_S(k_{say} \mid k) P(k)} \tag{6}$$

We ground out this recursive definition in a level 0 "literal" judge, who interprets the sender's message according to the literal semantics (Goodman & Frank, 2016). The literal judge's estimate of the truth directly matches their received message.

Results

Do Messages Become Increasingly Decoupled from Reality?

Probabilistic models serve as tools that help explain how emergent properties arise from the interactive dynamics of simple agents. Here, we examine how the properties of the communication channel change as a function of the senders' motives. We first tested how deceptive and cooperative senders adjust their messages to the predicted responses of judges over progressive levels of recursive reasoning. We hypothesize that the communication channel yields one of two potential patterns of stability. (1) Lies and truth inferences are amplified with each level of recursion, becoming increasingly decoupled from reality, and ultimately yield an unstable communication channel. Or (2) lies and truth inferences are checked by constraints of agents' goals, and ultimately converge on a stable, equilibrium state.

The simulation is initiated with the literal judge (Level 0, or L0) who directly estimates k_{est} from k_{say}. Next, a Level N (LN) cooperative or deceptive sender probabilistically decides what k_{say} (conditioned on k) to produce under the assumption that the judge behaves like an $N-1$ thinker. Then, an LN judge probabilistically decides what k_{est} (conditioned on k_{say}) to guess under the assumption that the LN sender behaves rationally. Senders and judges recursively reason in this way on and on. This method of modeling increasing depths of recursion is known in the economics literature as cognitive hierarchy or level-k reasoning (e.g. Camerer et al., 2004; Crawford and Iriberri, 2007; Cai and Wang, 2006). Figure 3 shows what senders message conditioned on the truth (green), what judges estimated about the truth conditioned on what message they received (red), and what judges estimated about the truth conditioned on the actual truth (orange).

As a basic validation, we show that cooperative senders say honest, unbiased messages, as observed by the shaded region along the identity line ($\Delta_{k_{say}-k} = -6.7 \times 10^{-19} \approx 0$; Fig. 3). The noise around what cooperative senders say arises from the probabilistic nature of the agents' decision rules. In turn, judges who expect the sender to be cooperative tend to interpret messages literally ($\Delta_{k_{est}-k_{say}} \approx 0$) and their resulting truth inferences are unbiased ($\Delta_{k_{est}-k} \approx 0$).

In contrast, deceptive senders say dishonest, biased messages ($\Delta_{k_{say}-k} = 0.28$). Critically judges who expect the sender to be deceptive correct for that bias in their interpretation of messages ($\Delta_{k_{est}-k_{say}} = -0.24$). Ultimately judges' resulting truth inferences are substantially less biased ($\Delta_{k_{est}-k} = 0.006$) than the messages and estimate, even if their estimates are noisier ($R^2 = 0.61$) than those of judges paired with cooperative senders ($R^2 = 0.74$). Importantly, even with greater recursion levels, messages converge to a stable equilibrium state, rather than becoming ever-increasingly decoupled from reality. We have shown convergence to this equilibrium state using simulation, without needing to formally derive Nash equilibrium.

How Do Senders' Motives Determine Equilibrium Form?

We propose that the sender's relative motive to have the judge mis-accurately infer the truth to the cost function to produce more extreme lies, or m, influences the sender-judge equilibrium state. We simulated 111 unique deceptive senders and varied the value of m between $\frac{1}{3}$ and 10. To ensure that agent pairs converged on an equilibrium state, we examined pairings at the (arbitrarily chosen) 20^{th} level of recursive reasoning. As m increases, we find that, at equilibrium, deceptive senders trend towards being more dishonest and thus produce a larger bias (Fig. 4). So too does the judge's bias correction

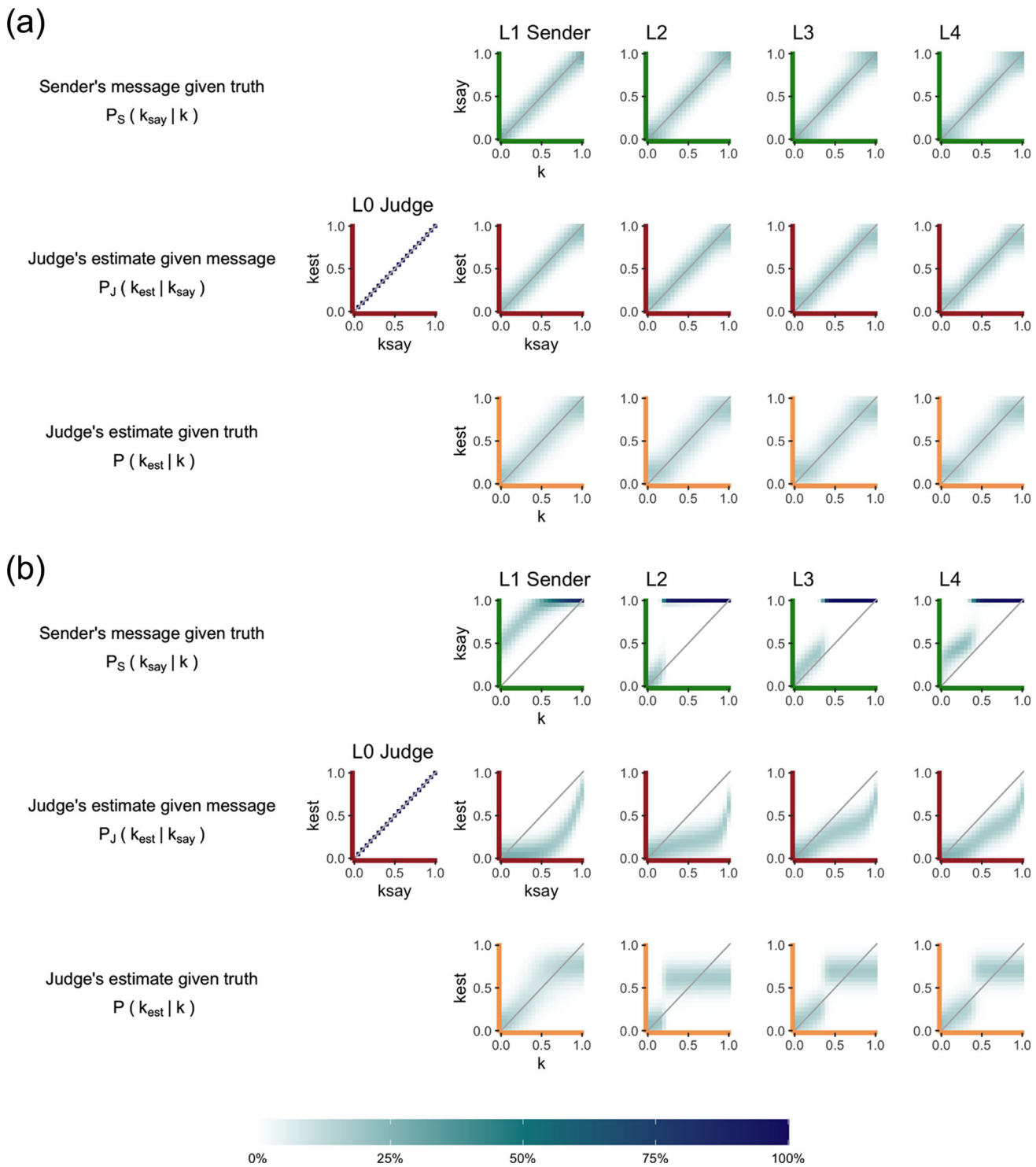

Fig. 3 Simulated behavior of sender-judge dyads over evolving levels of social reasoning (Levels L0 to L4). The shading reflects the probability the agent performs a behavior given the observation: green plots shows the sender's message conditioned on the truth, or $P_S(k_{say}|k)$; red shows the judge's estimate conditioned on the message they received, or $P_J(k_{est}|k_{say})$; and orange shows the judge's inference about the truth, or $P(k_{est}|k)$. (**a**) Cooperative senders produce unbiased messages, and judges' inferences about the truth are unbiased and have little noise. (**b**) Deceptive senders quickly converge on systematically producing biased messages. Judges' inferences about the truth are noisy but unbiased because they consider the sender's motive to deceive. The ratio of intended bias to message cost (m) is set to 1

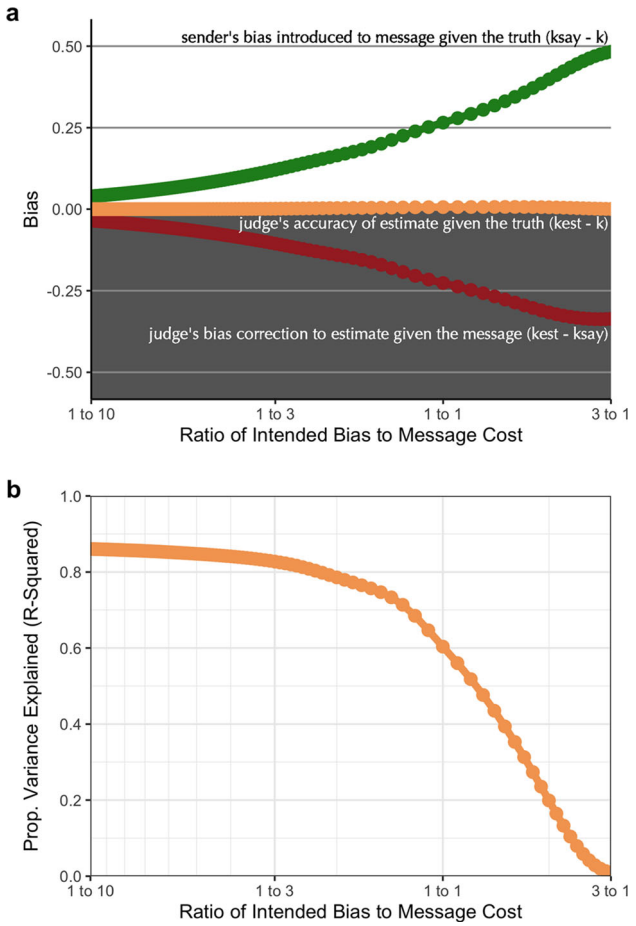

Fig. 4 Model's predicted bias and precision as a function of the ratio m. The x-axis is log scaled, with lower values representing when message cost dominates intended bias. (**a**) Senders' biases in their message relative to the truth (green; $k_{say} - k$) and judges' bias correction in their estimate relative to the message (red; $k_{est} - k_{say}$) both increase absolutely at higher ratios m, when intended bias dominates message cost (get further from 0). Regardless, the accuracy of judges' truth inference relative to the ground truth (orange; $k_{est} - k$) stabilizes at 0, implying that the truth gets unbiasedly conveyed to the listener in these communication systems even when messages are lies. (**b**) The proportion of variance explained (R^2) by the truth in judges' truth inferences decreases with higher ratios, meaning that there is less precision in communication systems in which speakers face relatively lower costs to lie

in the opposite direction. These initial results replicate previous findings from Kartik (2009), which showed that senders' incentives and costs influence what messages are said and how those messages are interpreted at equilibrium. Then, critically, we ask how m impacts judges' inferences about the truth ($k_{est} - k$) at equilibrium by testing how (1) accurate and (2) precise are judges' estimate of the truth relative to the ground truth.

We may expect accurate, or unbiased, truth inferences, in which case judges at equilibrium can perfectly correct for the bias introduced by senders. Alternatively, judges may *under*correct for senders' bias, so senders "win" in the long run because they succeed at causing the judge to overestimate. Or judges may *over*correct for senders' bias, so senders incidentally self-sabotage by leading judges to underestimate (counter to senders' goals). Each of these predictions appraise the success of the communication channel: do judges accurately extract the truth from communication, or do deceptive senders succeed at distorting judges' beliefs?

The model finds that judges' truth inferences are unbiased across all ratios m. Senders ultimately fail to distort judges' beliefs. This surprising finding calls into question the benefit to deceivers to lie when their motives are broadly suspected. A related configuration of deceptive communication channels is that while they may not bias the judges' inferences, the process may result in increased imprecision. For example, even if truth inferences were unbiased across repeated interactions, judges could still be inaccurate for most individual interactions.

How does precision in judges' truth inferences change as a function of the ratio m? We measured R^2, or the proportion of estimates' variance explained by the truth. Larger R^2 implies that judges' truth inferences are more consistent with the ground truth, while smaller implies that they are more distributed. The model finds that as the ratio of m increases, R^2 decreases to 0. In other words, truth inferences are more distributed when senders face a relatively lower cost to lie.

In sum, we found that deceptive senders' relative intended bias to distort judges' beliefs versus their message cost to produce larger lies drives the form of equilibria. In particular, we found that bias for both senders' messages and judges' inferences increases as intended bias increasingly dominates cost. Furthermore, we found that while accuracy in judges' truth inference is unbiased, precision decreases as intended bias increasingly dominates cost. Thus in communication channels where speakers face fewer costs to lying and judges suspect this, messages become more divorced from reality; nonetheless people extract accurate, albeit noisier, information about the truth.

How Do Judges' Bias Corrections Influence Speakers with Different Motivations?

Populations are composed of agents with varying motivations. While some speakers may be deceptive, more often speakers aim to be cooperative (Grice, 1975). How may speakers indirectly influence one another via socially reasoning about judges' behavior? In our model, speakers only directly interact with judges, not other speakers. This design allows us to isolate how speakers indirectly influence other speakers by way of judges' beliefs and actions. Specifically, we explore how cooperative speakers produce messages when they think the judge is correcting for deceptive speakers in the population?

Previous claims by Kartik (2009) state that speakers that are motivated to appear of a "higher" type should inflate their messages because judges, expecting inflated messages, deflate their inference and incidentally disadvantage speakers who lie less or are honest. Underlying assumptions of this speaker type is that they are adversarial and they are systematically penalized for judge underestimations. Yet in real world settings like letters of recommendation, letter writers more often think of themselves as helpful and cooperative. We expand on this previous claim to examine how even cooperative speakers, who want to guide listeners to accurate interpretations, may be driven to lie as well.

We examined the behavior of a cooperative L2 sender under the L1 judge's assumption about a mixed population of deceptive and cooperative senders. We varied the proportion of deceivers in the population, which scales how much bias the L1 judge assumes in the message and therefore how much they bias correct. The L1 judge's expected estimate of the truth gets fed to an L2 cooperative sender. Critically, to help the judge accurately infer the truth, the L2 cooperative sender adjusts their behavior to produce a *dishonest* message. Figure 5a shows an example simulation when the population is 50% deceptive.

We found that as the proportion of deceivers in the population increased, the L2 cooperative sender produced more bias in their messages (Fig. 5b). At the upper limit, when the cooperative sender expects the judge to believe the population is composed of 100% deceivers, the cooperative sender produces a bias of 0.24, which is still less than a deceptive sender with the same belief about the population, who produces a bias of 0.31. While the cooperative sender still biases their message, they do not do so to the extent of their deceptive counterpart.

Discussion

Detecting lies is often portrayed as a categorization process – is a message true or false? However, in many real world situ-

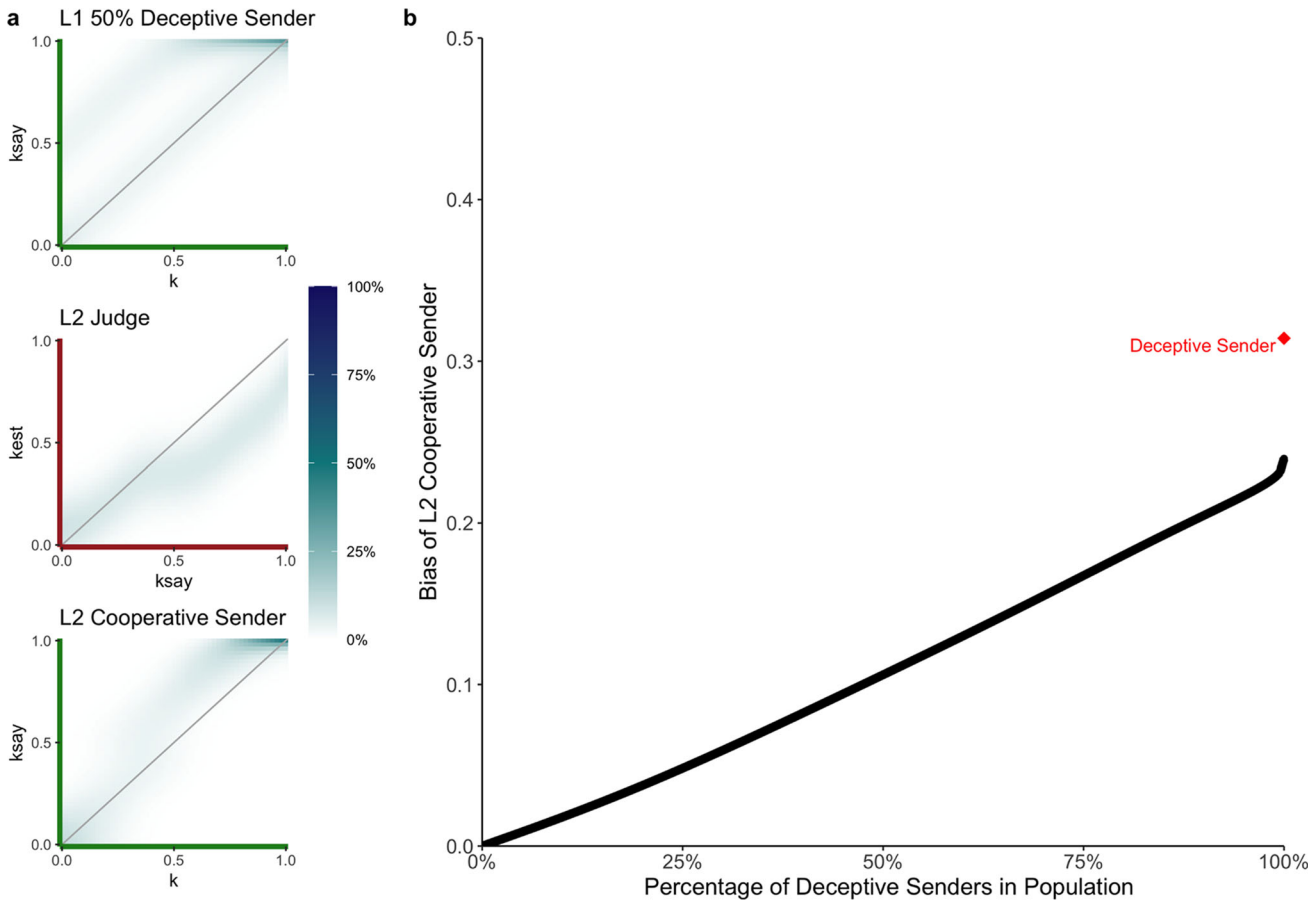

Fig. 5 (**a**) Simulated behavior of a cooperative sender who assumes the judge believes the population of senders is 50% deceptive and 50% cooperative. The top panel shows the bias of a mixture of messages from cooperative (on the identity line) and deceptive (off the identity line) L1 senders. The middle panel shows the L1 judges' bias correction for this mixture. The bottom panel shows how a cooperative L2 sender would bias their message (instead of being honest) to cater to the judge's bias correction. (**b**) The bias of L2 cooperative senders' messages increases with higher percentages of deceptive senders in the population. At 100% deceptive populations, L2 cooperative senders bias their messages less than L2 deceptive senders (red rhombus)

ations, people go beyond simply categorizing to make richer inferences – what is the actual ground truth? We focus on one goal in deception – that a speaker wants the judge to misestimate the truth (e.g. my candidate is the best fit for your position), while a judge's goal is to infer the truth (e.g. your candidate is an okay fit). Altogether, our studies show that a rational theory-of-mind framework explains how people may infer the truth from literally false messages. Behaviorally, we show that people can and do generate inferences about the truth from suspected lies, and they tune these inferences to their beliefs about senders' motives and costs. When speakers and listeners have veridical representations of each others' adversarial motivations and costs, the result is a state in which speakers say literally false messages but listeners nonetheless extract the truth. Probabilistic modeling shows speakers do not ratchet up to produce more and more extreme lies with increasing levels of theory-of-mind reasoning; instead, the communication channel stabilizes to an equilibrium state. For broad classes of systems, social reasoning about others predicts how accurately and precisely communication channels (e.g. letters of recommendation) transmit information about the ground truth. For individuals within systems, speakers' different motivations indirectly affect what others say, so that even cooperative speakers should be dishonest when they suspect listeners are correcting for deceptive speakers (e.g. cooperative recommenders should embellish what they say to accurately convey their belief to the reader).

While rational and recursive reasoning form the framework to explain how people infer truth from lies, two critical components serve as a precondition for such a system to get off the ground, that: (1) listeners know of speakers' directional deception goals and (2) bigger lies are more costly. Throughout this paper, we highlighted letters of recommendation as a communication system which, in equilibrium, messages are biased – recommenders inflate how positively they write about their candidate – yet the transmission is unbiased – readers extract accurate beliefs about the candidate. We speculate that these critical components coexist within many real-world communication systems, and thus our unified framework can explain idiosyncratic behaviors in communication systems that have not been linked previously. For example, when communicating your preferred political candidate via voting in run-off elections, voters can be honest by selecting their favorite candidate, or "strategic" by misrepresenting their preference in earlier rounds (Piketty, 2000). Both voting methods converge to different equilibria states – one is honest and one is dishonest, yet both result in the overall preferred candidate being elected. In essence, honesty and dishonesty are equally serviceable solutions to transmitting information. Then, there is puffery in marketing, which may not be perceived as false advertising because listeners make the adjustment to how they interpret the message (Stern & Callister, 2020). Even in communication systems that are not standardly thought of as deceptive, these principles may be applied to understand why populations form norms to produce nonliteral messages, such as in everyday hyperbole.

Until now, we have treated deceivers and cooperators as distinct agents. We defined cooperators as sharing the same goal as judges: to induce the judge to form an accurate belief about the world. Formally, we characterized deceivers as senders who have weighted incentives to induce a (mis)belief in the listener, but face weighted costs to saying more extreme lies. Cooperators can be mapped onto this formalism as well. Cooperators want to reduce the error in the accuracy of judge's estimates of the truth, juxtaposing deceivers that want to induce error. In this paper, we highlighted a cooperator that places zero weight on how they deviate their message from reality, although in principle cooperative senders may prefer to be honest, as deceivers do. Thus, our framework presents a unifying factor between cooperators and deceivers in their motive to influence the listener's beliefs.

The cost of lying plays a critical role in driving how communication appears in these distorted communication systems. Indeed, intrinsic aversions to lying have helped to explain why people tell the truth or overcommunicate information when being completely uninformative is in fact the theoretically optimal solution (e.g. Hurkens and Kartik, 2009; Cai and Wang, 2006). The cost of lying also plays a critical role in driving how communication appears in these distorted communication systems. Our behavioral experiment showed that people can make reasonable inferences about the truth from what they expect about speakers' costs to lie. Our choice in manipulating physical costs was intended to make the manipulation experimentally tractable. Real world costs faced by speakers are more cognitive in nature, such as thinking up a large yet still plausible lie or inhibiting one's intrinsic and moral aversions to lying. Within psychological research, physical and cognitive effort have long been viewed as analogous, though not perfectly one-to-one, subjective experiences (Eisenberger, 1992; Kool et al., 2010). Whether the cost is physical or cognitive, we expect speakers to adjust their lies based on these costs and judges to calibrate their inferences to changes in costs across contexts. However, there are limitations to operationalizing cost as physical effort, as we do in the experiments. Focusing on physical costs implicitly overlooks additional costs that participants instinctually feel. Additionally, the subjective cost may not scale linearly with the objective physical cost. Costs might also scale variably across people. Our behavioral results may have an inflated sense of how accurately people can tune into their opponents' m, whereas in the real world people may have a noisier representation of others' m without direct access to contexts' effect on cognitive costs.

Our model showed that as cost functions decrease, messages are more dishonest. Cost drives communication systems to approach an equilibrium state that preserves accurate inferences about the truth, even when speakers are deceptive. Formally, the emergence of equilibrium depends on a crossover effect between the linear incentive to distort listeners' belief versus the quadratic cost to produce larger lies. When this crossover effect is no longer valid – in our formalization, this happens when the incentive to distort listeners' belief far surpasses the cost – the communication system takes on a ratcheting effect, in which lies become increasingly extreme. When lies are so extreme that they become decoupled from reality, listeners no longer extract signal from the message, so accurate inferences to the truth end up as an incidental byproduct of listeners randomly guessing what the truth could be. These "runaway lies" point to the value of intensifying liars' costs to producing more extreme lies. Whether lying costs are increased via interventions targeting individuals' cognitive load or reputational risks, or by improving detection algorithms, listeners would gather more signal from lies and more precisely infer the truth.

Now that we have characterized communication systems that advantage listeners, conversely we can better understand when listeners' inferences go awry. Listeners may fall prey to deceptive speakers if they have an incorrect model of their opponent. In general, people are boundedly rational agents who face computational constraints (Lieder & Griffiths, 2020), such as limited recursive, or level-k, reasoning (Camerer et al., 2004; Crawford & Iriberri, 2007; Stahl, 1993; Kawagoe & Takizawa, 2009; Wang et al., 2010). Therefore, accurate opponent modeling is not only a challenge because individuals are bounded, but their opponents are as well. Additionally, while people may be aware of broad goals within a given communication system (e.g. letter writers generally want to promote their candidate), they may not be fine-tuned to individuals' motives and costs. For example, in general people would not suspect that a letter is downright fabricated, assuming that most people face higher costs to drastic deceptions. Therefore, individual fabricators benefit from readers who under-correct the bias by assuming that the letter is embellished, but not that it is fabricated. Meta-reasoning deceivers may even actively conceal their motives and costs. Of course, speakers who fake a cooperative intention build trust with the listener and are the most successful deceivers. But an even richer (untested) prediction from this work is that speakers, even when transparent about their deceptive intent, can conceal how *strong* their intention is to deceive to *gradedly* dupe their listener.

In conclusion, people's intuitive theory-of-mind reasoning – and not necessarily the assumption that others are cooperative – allows listeners to infer the truth from literally false messages, so long as they are equipped with sufficient knowledge about the speakers' goals. Taking a first principles approach to agents' goals, costs, and actions, we bridged individual listeners to broad classes of distorted communication systems, to characterize how they both systematically transmit and interpret information. Lastly, these results call into question the traditional depiction of people as naïve lie detectors, and instead support a nuanced depiction of people as robust lie interpreters.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s42113-023-00187-0.

Acknowledgements We would like to thank Frank Mollica, Judy Fan, Robert Hawkins, Lindsey Powell, Holly Huey, and Aarthi Popat for various discussions that contributed to the final product of this work. Lastly, we thank Chaz Firestone for inspiring this work with his Tweets.

Author Contributions All authors contributed to the conception and design. Material preparation, data collection, and analysis were performed by LAO. The manuscript was written by LAO and EV. All authors approved the final manuscript.

Funding This material is based upon work supported by the NSF Graduate Research Fellowship under Grant No. DGE-1650112 to LAO.

Data Availability Data from all participants are fully anonymized.

Code Availibility Data and code for the experiment and analysis are available at https://github.com/la-oey/InferringTruth

Declarations

Financial statement The authors have no relevant financial or non-financial interests to disclose.

Informed consent Informed consent was obtained from all human participants, and the study was approved by the university's Institutional Review Board.

Conflicts of interest The authors declare no conflict of interest.

References

Abeler, J., Nosenzo, D., & Raymond, C. (2019). Preferences for truth-telling. *Econometrica, 87*(4), 1115–1153.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires, and percepts in human mentalizing. *Nature Human Behaviour, 1*(4), 1–10.

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition, 120*(3), 322–330.

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review, 10*(3), 214–234.

Bond, C. F., & DePaulo, B. M. (2008). Individual differences in judging deception: Accuracy and bias. *Psychological Bulletin, 134*(4), 477–492.

Cai, H., & Wang, J.T.-Y. (2006). Overcommunication in strategic information transmission games. *Games and Economic Behavior, 56*(1), 7–36.

Camerer, C. F., Ho, T.-H., & Chong, J.-K. (2004). A cognitive hierarchy model of games. *Quarterly Journal of Economics, 119*(3), 861–898.

Clark, H. H. (1996). *Using Language*. Cambridge, UK: Cambridge University Press.

Crawford, V. P., & Iriberri, N. (2007). Level-k auctions: Can a nonequilibrium model of strategic thinking explain the winner's curse and overbidding in private-value auctions? *Econometrica, 75*(6), 1721–1770.

Crawford, V. P., & Sobel, J. (1982). Strategic information transmission. *Econometrica*, 1431–1451

Debey, E., De Houwer, J., & Verschuere, B. (2014). Lying relies on the truth. *Cognition, 132*(3), 324–334.

Eisenberger, R. (1992). Learned industriousness. *Psychological Review, 99*(2), 248–267.

Feiler, D. C., Tong, J. D., & Larrick, R. P. (2013). Biased judgment in censored environments. *Management Science, 59*(3), 573–591.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science, 336*, 998.

Franke, M., Dulcinati, G., & Pouscoulous, N. (2020). Strategies of deception: Under-informativity, uninformativity, and lies – misleading with different kinds of implicature. *Topics in Cognitive Science, 12*(2), 583–607.

Gerlach, P., Teodorescu, K., & Hertwig, R. (2019). The truth about lies: A meta-analysis on dishonest behavior. *Psychological Bulletin, 145*(1), 1–44.

Gneezy, U. (2005). Deception: The role of consequences. *American Economic Review, 95*(1), 384–394.

Gneezy, U., Kajackaite, A., & Sobel, J. (2018). Lying aversion and the size of the lie. *American Economic Review, 108*(2), 419–453.

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

Grice, H. P. (1975). Logic and conversation. Speech ActsIn P. Cole & J. L. Morgan (Eds.), *Syntax and Semantics* (Vol. 3, pp. 64–75). New York: Academic Press.

Gweon, H., Pelton, H., Konopka, J. A., & Schulz, L. E. (2014). Sins of omission: Children selectively explore when teachers are underinformative. *Cognition, 132*(3), 335–341.

Hayes, B. K., Banner, S., Forrester, S., & Navarro, D. J. (2019). Selective sampling and inductive inference: Drawing inferences based on observed and missing evidence. *Cognitive Psychology, 113*, 101221.

Hogarth, R. M., Lejarraga, T., & Soyer, E. (2015). The two settings of kind and wicked learning environments. *Current Directions in Psychological Science, 24*(5), 379–385.

Hurkens, S., & Kartik, N. (2009). Would I lie to you? On social preferences and lying aversion. *Experimental Economics, 12*, 180–192.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences, 20*(10), 589–604.

Kartik, N. (2009). Strategic communication with lying costs. *Review of Economic Studies, 76*(4), 1359–1395.

Kawagoe, T., & Takizawa, H. (2009). Equilibrium refinement vs. level-k analysis: An experimental study of cheap-talk games with private information. *Games and Economic Behavior, 66*(1), 238–255.

Kool, W., McGuire, J. T., Rosen, Z. B., & Botvinick, M. M. (2010). Decision making and the avoidance of cognitive demand. *Journal of Experimental Psychology: General, 139*(4), 665–682.

Leach, A.-M., Talwar, V., Lee, K., Bala, N., & Lindsay, R. C. L. (2004). "intuitive" lie detection of children's deception by law enforcement officials and university students. *Law and Human Behavior, 26*(6), 661–685.

Levine, T. R., Park, H. S., & McCornack, S. A. (1999). Accuracy in detecting truths and lies: Documenting the "veracity effect". *Communication Monographs, 66*(2), 125–144.

Lewis, D. (1969). *Convention*. Harvard University Press.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences, 43*, e1.

Lundquist, T., Ellingsen, T., Gribbe, E., & Johannesson, M. (2009). The aversion to lying. *Journal of Economic Behavior & Organization, 70*(1–2), 81–92.

Maggian, V., & Villeval, M. C. (2016). Social preferences and lying aversion in children. *Experimental Economics, 19*, 663–685.

Oey, L. A., Schachner, A., & Vul, E. (2023). Designing and detecting lies by reasoning about other agents. *Journal of Experimental Psychology: General, 152*(2), 346–362.

Piketty, T. (2000). Voting as communicating. *Review of Economic Studies, 67*(1), 169–191.

Pinker, S., Nowak, M. A., & Lee, J. J. (2008). The logic of indirect speech. *Proceedings of the National Academy of Sciences, 105*(3), 833–838.

Ransom, K., Voorspoels, W., Navarro, D. J., & Perfors, A. (2019). Where the truth lies: How sampling implications drive deception without lying. *PsyArXiv*

Schelling, T. C. (1960). *The strategy of conflict*. Cambridge, MA: Harvard University.

Shalvi, S., Handgraaf, M. J. J., & De Dreu, C. K. W. (2011). Ethical manoeuvring: Why people avoid both major and minor lies. *British Journal of Management, 22*, 16–27.

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

Stahl, D. O. (1993). Evolution of smartn players. *Games and Economic Behavior, 5*(4), 604–617.

Stern, L. A., & Callister, M. (2020). Exploring variations of hyperbole and puffery in advertising. *Journal of Current Issues and Research in Advertising, 41*(1), 71–87.

ten Brinke, L., Vohs, K. D., & Carney, D. R. (2016). Can ordinary people detect deception after all? *Trends in Cognitive Sciences, 20*(8), 579–588.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences, 28*(5), 675–691.

van't Veer, A. E., Stel, M., & van Beest, I. (2014). Limited capacity to lie: Cognitive load interferes with being dishonest. *Judgment and Decision Making, 9*(3), 199–206.

Walczyk, J. J., Harris, L. L., Duck, T. K., & Mulay, D. (2014). A social-cognitive framework for understanding serious lies: Activation-decision-construction-action theory. *New Ideas in Psychology, 34*, 22–36.

Wang, J.T.-Y., Spezio, M., & Camerer, C. F. (2010). Pinocchio's pupil: Using eyetracking and pupil dilation to understand truth telling and deception in sender-receiver games. *American Economic Review, 100*(3), 984–1007.