Lies are crafted to the audience

Lauren A. Oey (loey@ucsd.edu) and Edward Vul (evul@ucsd.edu)
Department of Psychology, University of California, San Diego
9500 Gilman Dr., La Jolla, CA 92093 USA

Abstract

Do people cater their lies to their own beliefs or others' beliefs? One dominant *individual-based* account considers lying to be an internal tradeoff between self-interest, norms, and morals. However, recent *audience-based* accounts suggests that lying behavior can be better explained within a communicative framework, wherein speakers consider others' beliefs to design plausible lies—highlighting the role of theory-of-mind in strategic lying. We tease apart these accounts by examining human lying behavior in a novel asymmetric, dyadic lying game in which speakers' beliefs differ from those they ascribe to their audience. We compare participants' average reported lie (controlling for the truth) across conditions that manipulated the player's and the audience's beliefs. We find that people spontaneously tune their lies to beliefs unique to their audience, more than to their own beliefs. These results support the audience-based account of lying: estimates of how listeners will respond determine how people decide to lie.

Keywords: deception; partial observability; theory of mind

Introduction

Human communication is generally honest—people speak the truth, and they assume others do as well (Abeler, Nosenzo, & Raymond, 2019; Levine, 2014)—but this expectation of honesty renders listeners susceptible to deceptive speech. More generally, listeners' expectations about reality determine which statements are likely to be seen as lies. Do liars spontaneously design lies in accordance with their estimates of their audience's beliefs? The ability to lie at all seems to require theory-of-mind (ToM), or the ability to reason about others' mental states (e.g. Ding, Wellman, Wang, Fu, & Lee, 2015), suggesting that representing the beliefs of the audience is critical to lying. After all, intentional lying is predicated on the understanding that the audience might form beliefs different from the speaker. But other research has called into question whether this level of ToM reasoning is necessary to achieve human-like lying behavior.

One popular account proposes people's lies are largely constrained by their own beliefs, e.g. about around honesty (Gino, Ayal, & Ariely, 2009) and morals (Mazar, Amir, & Ariely, 2008). This *individual-based* account was designed specifically to explain why people have a tendency to lie but avoid large lies, even in situations without a risk of detection. Invoking beliefs about morals, norms, and self-conception has influenced policy-making: rather than policies focused on detecting and punishing wrongdoers, a seemingly easier method nudges people to behave honestly (but see Verschuere et al., 2018; Kristal et al., 2020 for failures to replicate prominent intervention experiments). Under the individual-based account, people's lies should be driven by their own values and prior beliefs about the world. Speakers will avoid lies that seem big to them, even if to their audience that lie would be small, and undetectable, and vice versa.

Furthermore, a resource rational argument can be made for why catering to audience-specific beliefs may be uneconomical, even when at risk of detection. First, people are practically at chance when detecting lies, as shown in experimental studies (Bond & DePaulo, 2006), and this mediocre performance is in part attributed to people's attention to ineffective cues (Vrij, Granhag, & Porter, 2010). Under a blanket assumption that detectors are simply guessing, a liar need not attribute sophisticated reasoning to their audience to succeed in duping them. Second, lying is effortful—it is cognitively demanding (Vrij, Fisher, Mann, & Leal, 2006) and incurs longer response times than telling the truth, even without the risk of getting caught (Suchotzki, Verschuere, Van Bockstaele, Ben-Shakhar, & Crombez, 2017; Capraro, Schulz, & Rand, 2019, but see Shalvi, Eldar, & Bereby-Meyer, 2012). Applying a complex ToM process to reason about the audience would presumably add to the cognitive demand required of lying (Apperly, Riggs, Simpson, Chiavarino, & Samson, 2006). Finally, given the scarcity of distinguishing information about others' idiosyncratic beliefs, a heuristic that relies only on the speakers' own beliefs to choose a lie may be a globally optimal heuristic.

In contrast to these individual-based accounts, some recent work proposes that ToM is necessary to explain how people lie when an audience has the opportunity to detect lies, as it is used to predict the listeners' likely response. These accounts are based on traditional economic approaches to why people commit crimes in adversarial situations (Becker, 1968). Such *audience-based* accounts formalize the decision processes underlying lying (and lie detection) within a communicative framework: when sending a message, people consider others' beliefs to design plausible lies; when receiving that message, people consider others' goals to discriminate likely from unlikely lies (Oey, Schachner, & Vul, 2019). Similar communicative frameworks have found success in explaining human preferences for various strategies of deception across contexts (e.g. Montague, Navarro, Perfors, Warner, & Shafto, 2011). For example, audiences' coopera-

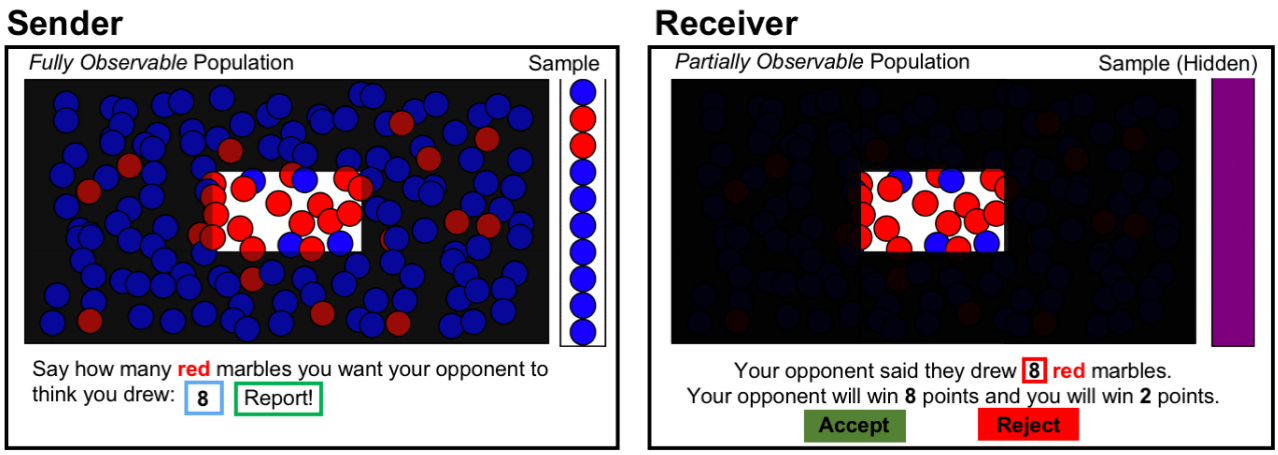

Figure 1: Design of the game. For the sender, beliefs about the base-rate are fully observable: the sender knows the distribution of red and blue marbles observed by the receiver (in the inner white box), and the overall distribution (in the inner white and surrounding black box). For the receiver, beliefs about the base-rate are partially observable: the receiver can only observe the distribution of marbles in the window (the inner white box). Here, the sender believes the full population contains 20% red and 80% blue marbles, and they know the receiver observes a subset of the population that is 80% red and 20% blue marbles.

tive expectations drive deceivers to elect to be uninformative over the alternative choice to mislead their audience to a false conclusion (Ransom, Voorspoels, Navarro, & Perfors, 2019; Franke, Dulcinati, & Pouscoulous, 2020). Moreover, general uncertainty about the audience's intent leads people to build plausible deniability into their message by speaking indirectly, such as in bribing (Lee & Pinker, 2010). And prominently, lying is more prevalent when it benefits the expected payoff of the audience, as in white lie-telling (Gneezy, 2005; Erat & Gneezy, 2012).

Broadly, the individual- and audience-based accounts generate distinct predictions about how robust human lying and lie detecting behaviors are to beliefs about other agents' minds. Rigorously testing the audience-based account of lying requires showing that people cater their lies specifically to the opponent's prior beliefs. Alternatively, if people only cater their lies to their own prior beliefs, this evidence would favor an individual-based account of lying. One study, Oey et al. (2019), attempted to tease apart these accounts by manipulating *global* base-rate beliefs about the world to see if they systematically influence how people produce lies (see also Mazar et al., 2008). However, the manipulation perturbed shared information about base-rates, and so it could not tease apart whether senders are adjusting to *their own* or *their opponent's* beliefs about base-rates. To clearly demonstrate that people can and do tailor their lies to their audience, we must show that that the lies people tell vary based on what they think their audience believes. To our knowledge, such a study has yet to be described in the literature.

In this study, we aim to fill this gap by asking the nuanced question: do people tune the frequency and content of their lies based on expectations about the unique beliefs of their audience? We test this question in an asymmetric, dyadic lying game where speakers are led to believe that listeners have beliefs that differ from their own. If people anchor their lies based on their own belief instead of their audience's, this would suggest a cognitive limitation in people's lying ability—perhaps individual-based considerations are sufficient to characterize human lying. Alternatively, if people anchor their lies based on the audience's belief, this would support an audience-based account—people spontaneously consider the audience's response when designing their lies.

Experiment

Participants played a dyadic lying game, alternating between both roles (sender and receiver) between each trial. To control for the behavior of the opponent, participants played against an AI. Participants were instructed that their goal was to defeat their opponent by the most points possible. This experimental lying paradigm was inspired by the sender-receiver game used in Oey et al. (2019), with some key improvements.

Participants

291 participants were recruited from the undergraduate population at the University of California, San Diego to participate in an online study. Participants received course extra credit for their time. Of these, 33 were excluded for failing to sufficiently answer at least 75% of the attention check questions.

Design

In the game, the sender and receiver both observe a box containing some population of red and blue marbles. In private, the sender randomly samples 10 marbles out of the box ("the truth" is how many red marbles they sample), and then reports to the receiver how many red marbles they supposedly sampled (which can correspond to either the truth or any lie between 0 and 10). The receiver does not see the true sample, and they decide whether to accept or reject (i.e. accuse as a lie) the sender's report.

The sender is motivated to report more red marbles for a higher gain in points, but they are dissuaded from getting caught in a lie for a penalty. If the receiver accepts the sender's report, the sender earns points for the red marbles reported while the receiver earn points for the blue marbles reported (ten minus the red marbles reported). For example, if the receiver accepts when the sender reports that they sampled 7 red marbles, then the sender would receive 7 points, and the receiver 3, regardless of how many red marbles the sender actually saw. However, the receiver may choose to reject the report if they are suspicious. If the receiver rejects the report, and the report was actually a lie, then the receiver always gains 5 points while the sender loses 5. If the receiver rejects a report that was actually true, then the sender gains points for the number of red marbles seen and reported, and the receiver gains points for blue marbles and pays a penalty of 5 points for falsely accusing the sender of lying. Together these payoffs motivate the sender to lie, but not be caught, and the receiver to catch lies, but not make false accusations. After four practice trials, participants only received intermittent feedback (every fifth trial) about their gameplay in the form of both players' cumulative points.

To tease apart the audience- and the individual-based accounts of lying, we manipulated the distribution of red and blue marbles in the box visible to the sender and the receiver. The box contains a window on one side (an inner white box) through which the receiver can see the distribution of red and blue marbles. The other side is open—the sender can see what the receiver sees through the window (the inner white box), as well as the full distribution of red and blue marbles (the inner white box and the surrounding black box). In other words, the population of marbles is fully observable for the sender, but it is partially observable for the receiver. Furthermore, the sender can infer how the receiver's base-rate differs from their own, but the receiver has no information on which to evaluate whether the sender has a belief different from their own.

We used a 3×3 within-subject design: the sender's base-rate (total box) was 20%, 50%, or 80% red; the receiver's base-rate (inner white box) was 20%, 50%, or 80% red. These conditions were randomly sampled for each trial. Altogether, this set up made the receiver particularly susceptible to deception in certain conditions when the sender's and the receiver's beliefs are asymmetric.

A possible concern of this setup is that participants' *beliefs* about the base-rate may not actually correspond to the *veridical* base-rate of marbles (the raw counts of red and blue marbles in the box). For instance, a key assumption of the study is that senders' beliefs about the receiver's belief are different from their own. To test the soundness of our assumptions about players' beliefs, we asked participants to respond on a slider scale about the distribution of marbles from their own or their opponent's perspective (shown in Fig. 2). The left side of the slider bar was red and the right side was blue, so that the further rightward the bar was dragged, the more

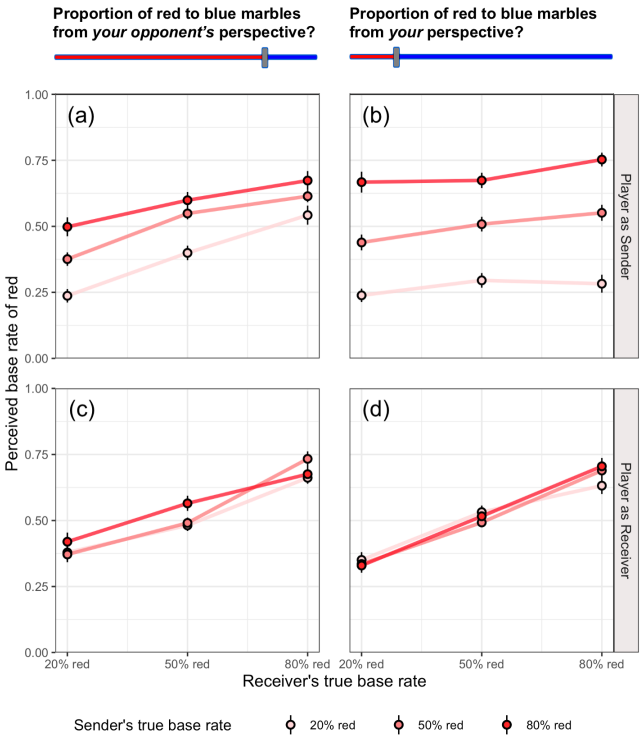

Figure 2: Participants were asked about the distribution of red/blue marbles from either their opponent's (a, c) or their own perspective (b, d). Their response was recorded using a slider scale. (a) Senders believed receivers' base-rate beliefs shifted with the receivers' true base-rate as expected, but surprisingly, the senders' true base-rate also had a small influence on their response. (b) Senders accurately assessed their own base-rate. Receivers responded the same for (d) their own perspective as (c) their opponent's perspective.

the bar was "filled in red." Labels below the slider ("more blue" to the left, "more red" to the right) helped to clarify the scale's direction. Participants were also intermittently asked attention check questions about how many red marbles were drawn or reported. The questions were randomly distributed throughout the experiment. All participants received a total of 19 base-rate and attention check questions, except three subjects who received 18. Participants played for a total of 100 trials.[1]

Results

Manipulation Check

Did our manipulation of sender's beliefs, receiver's beliefs, and sender's beliefs about the receiver have the intended effects? For our manipulation to work, we required that three conditions be satisfied. (1) The sender ought to recognize that the distribution of red and white marbles in the inner white box is visible to the receiver and guides the receiver's beliefs. (2) In inferring the receiver's beliefs, the sender must

[1]Data and code for experiment and analysis are available at https://github.com/la-oey/ConcealedLies

recognize that the players can hold different beliefs about the base-rate of marbles. (3) For the receiver to be susceptible to exploitation, the receiver needs to believe the base-rate they see. Finally, although not critical to our primary experimental goals, it is useful to ask (4) does the sender accurately assess the receiver's belief? We considered whether participants' base-rate estimates (ranging from 0 to 100) varied as expected with player role (sender or receiver), question type (own or opponent's belief), and sender and receiver base-rate conditions.

(1) Do senders notice receivers' base-rate? We checked if the study's key manipulation was successful—that senders were aware that they had access to what receivers could see through the box's window (Fig. 2a). A two-way ANOVA with an interaction revealed a significant effect of receiver base-rate ($F(6,636) = 17.06$, $p < 0.0001$), suggesting that senders understood that receivers' beliefs about the base-rate were constrained by the aperture. There was also a significant effect of sender base-rate ($F(6,636) = 11.59$, $p < 0.0001$), indicating some "leakage" of senders' beliefs into their assessment of receivers' beliefs.

(2) Do senders believe receivers have different beliefs from themselves? Our manipulations were specifically aimed to induce an asymmetry between senders' beliefs about receivers' beliefs (Fig. 2a) and the senders' own beliefs (Fig. 2b). We tested whether *whose* beliefs (sender or receiver) the sender was asked about interacted with the receiver and sender base-rate conditions, separately. For both interactions we found a significant effect (with receiver base-rate: $F(2,1242) = 10.72$, $p < 0.0001$; with sender base-rate: $F(2,1242) = 21.74$, $p < 0.0001$). This means that our manipulations succeeded at separately influencing the senders' estimates of the base-rate, and their assessments of the receivers' beliefs about that base-rate.

(3) Do receivers assume senders share the same beliefs as themselves? Another assumption of the study is that receivers assume that the distribution of marbles visible to them approximately matches the distribution of marbles from which the sender is sampling. As the receiver, the participant may distrust that the sender's distribution corresponds to the receiver's. In this case, the receiver may instead assume, in spite of the evidence for the distribution that they see, the sender is actually sampling from a noisy distribution centered at 50% red. We compared receivers' beliefs about senders' beliefs (Fig. 2c) and their own beliefs (Fig. 2d). We tested a model with an interaction between sender and receiver base-rate conditions and an additive effect of question type to predict receivers' responses. We found that question type—whether the receiver was asked about their own, or their opponent's beliefs—did not significantly improve the model ($F(1,1211) = 3.52$, $p = 0.06$). This corroborates our assumption that receivers default to the assumption that the senders' beliefs approximate their own.

(4) Do senders' beliefs about receivers' beliefs map onto receivers' own beliefs? While it is not necessary that senders accurately assess receivers' base-rate beliefs, it is clear from the data that sender base-rate has an additive effect on senders' assessment of receivers' beliefs (Fig. 2a), when compared to receivers' own beliefs (Fig. 2d). In other words, senders' beliefs about receivers' beliefs are not perfectly accurate. However, our findings do not rely on the senders' perfect assessment of receivers' beliefs.

In aggregate, our manipulations worked. Senders recognized that receivers' beliefs about the base-rate were different from their own, and receivers used the information visible to them to approximate senders' likely beliefs.

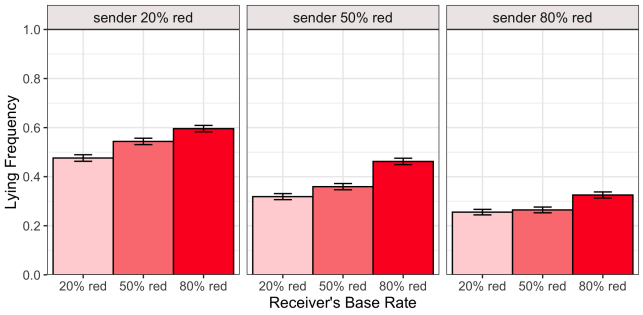

Figure 3: The rate of lying (as opposed to telling the truth) across conditions. There is as an effect of the receivers' (x) and the senders' base-rate condition (panels). People lie more when the receivers' base rate belief is higher (e.g. 80%), suggesting that people recognize when their audience is more exploitable.

Lies

The behavior of senders can be factored into (a) their rate of lying (as opposed to truth-telling) and (b) the lie they told when choosing to lie. Here, we define a lie as any reported value that was false, regardless of its intention. As we cannot pinpoint participants' underlying intentions, reports grouped into this category may have been intentional lies designed to advance the player in the game, accidental false reports, etc. We compute the rate of lying (a) as the proportion of false reports to all reports. The lie told (b) is the report, conditioned on the true number of red marbles sampled and the report being false.

When do people lie? Our first analysis aims to characterize the rate of lying. Broadly, in our data set, participants lied 40% of the time. We break this down by condition in Figure 3, where we show that senders' lying rates vary as a function of the true base-rate experienced by the sender ($\chi^2(2) = 665$, $p < 0.0001$) as well as the base-rate beliefs they attribute to the receiver ($\chi^2(2) = 138$, $p < 0.0001$).

A logistic mixed effect regression model including both sender and receiver base-rate beliefs revealed significant differences across receivers' base-rate conditions (50% vs 20% red: $\hat{\beta} = -0.23$, $z = -4.70$, $p < 0.0001$; 50% vs 80% red: $\hat{\beta} = 0.34$, $z = 6.99$, $p < 0.0001$). This means that people tended to lie more when receivers' beliefs about the base-rate were higher (e.g. 80% red)—circumstances in which

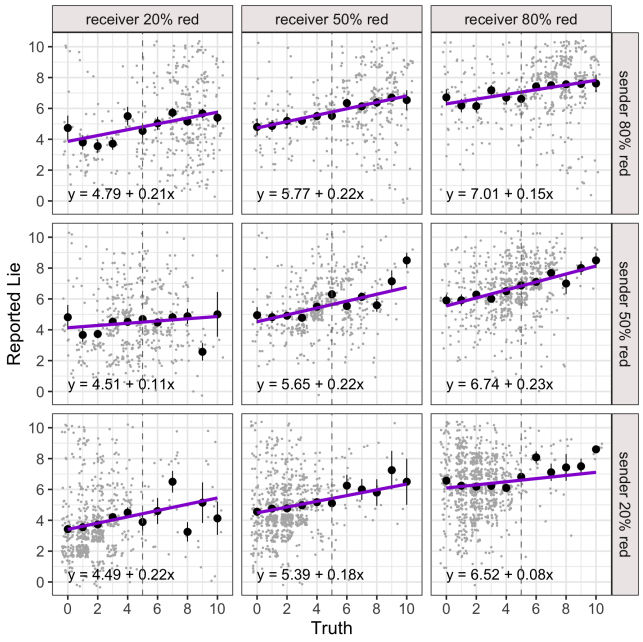

Figure 4: The distribution of lies across each condition. Each gray point was a false reported value. A linear mixed effect model was fit to each condition, with intercepts centered at Truth = 5. Intercepts were largely unchanged across different levels of sender's belief (columns), while they changed dramatically as a function of receiver's belief (rows). This pattern shows that the lies people tell in our game reflect mostly the receiver's base-rate belief.

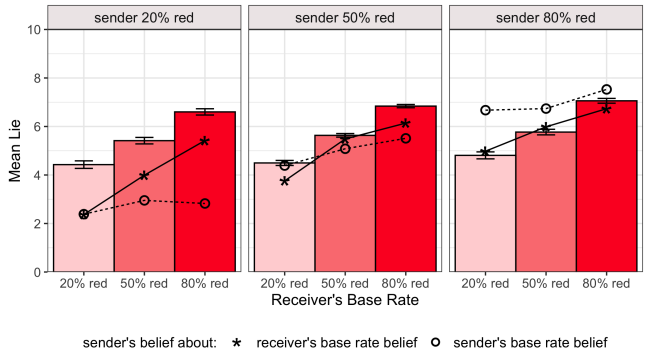

Figure 5: The average lie across conditions, computed from the intercept of the linear fit (from Fig. 4). There is a strong effect of the receivers' base-rate condition (x), and little effect of the senders' base-rate condition (panels). Star represents the senders' estimates of the receivers' belief about the base-rate (from Fig. 2a), and the circle represents the sender's direct estimate of the base-rate (from Fig. 2b). The average lie appears to closely track senders' estimates of receivers' beliefs, suggesting that senders use theory-of-mind to choose how to lie.

the sender may tell larger lies while maintaining plausibility. These results imply that people can recognize when their audience is more exploitable, and they take advantage of these situations.

How do people lie? In our second analysis, we focus on how people chose which lies to tell. We examined how the relationship between the truth (i.e. number of red marbles truly sampled) and reported lies (i.e. reported red marbles when those differed from the truth) varied across sender base-rate and receiver base-rate conditions (Fig. 4). We fitted a linear regression to the number of marbles falsely reported as a function of a three-way interaction between the truth and the senders' and the receivers' base-rate. Subject was included as a random intercept. To facilitate comparisons across conditions, the truth values were centered on 5 so that the models' intercepts correspond to the lies told when 5 marbles were truly drawn. Thus, changes in the intercept reflect changes in which lies are likely to be told in response to seeing 5 red marbles actually drawn.

First, we examined the general relationship between what the speaker saw, and what lie they reported. As expected, people falsely reported larger numbers when they drew more marbles in reality ($\hat{\beta} = 0.21$, $t(5150) = 18.83$, $p < 0.0001$, $r = 0.25$). After all, it does not make sense to falsely report fewer marbles than were actually seen, so if someone tells a lie in response to a large number, it is likely to be a large lie, in absolute terms.

This positive relationship also relates to the distance between the lie and the truth as a function of what the truth is. Although there are good theoretical reasons to think about this relationship, our task and data cannot meaningfully speak to these because the results reflect largely the range of possible responses. If the goal of the sender is to over-report how many red marbles they saw, then when the reality is that fewer red marbles were sampled, there is a greater margin for over-reporting. This means that people cannot possibly lie by the same magnitude when they see a large number as compared to when they see a small number. In Fig. 4, a slope of 1 would indicate a constant difference between truth and lies regardless of how many red marbles were actually drawn. A slope of 1 for these task results would be impossible unless the average lie magnitude was 0—when the truth was 10, speakers cannot possibly lie in the positive direction, since they can only report numbers between 0 and 10. Our results showed a much shallower slope of 0.21, revealing that the magnitude of the lie was smaller for larger truths. However, because this result is inevitable given the structure of the task, we cannot say whether the apparent behavior arises because people are less inclined to tell big lies when the reality is already in their favor. Such questions would require a different task without a restricted reporting range.

Next, critically for our main question, we analyzed whether the base-rate conditions influenced people's lies. Both of the base-rate conditions were significant predictors of the reported lies, but the receivers' base-rate had a greater effect on lies ($\chi^2(12) = 1214.7$, $p < 0.0001$, $\hat{\omega}^2 = 0.119$) than the senders' base-rate ($\chi^2(12) = 34.7$, $p = 0.0005$, $\hat{\omega}^2 = 0.003$). Thus, senders weighed receivers' prior beliefs *more* than their own when deciding how to lie. These results point to peo-

ple's abilities to construct gain-increasing lies around the audience's unique beliefs.

To further elucidate how beliefs influence senders' lying behavior, we compared the elicited senders' beliefs (about their own and their opponents' base-rate beliefs; Fig. 2) against the average lie (i.e. intercepts in Fig. 4). This comparison shows that the average lie appears to track senders' beliefs about receivers' beliefs more than senders' own beliefs (Fig. 5). These results support the claim that senders are using an audience-based strategy to choose their lies.

Discussion

When, and how do people lie? The currently dominant view considers lying to be an internal tradeoff between self-interest, self-conception, self-serving justification, norms, and morals (Jacobsen, Fosgaard, & Pascual-Ezama, 2018). This view suggests that lies are limited by individuals' desire to be virtuous, and small lies are seen as smaller sins than larger lies. This account seems to explain why people tend to lie rarely, and why they avoid large lies; even when they are in situations with minimal risk for getting caught (e.g. Mazar et al., 2008). Meanwhile, a growing area of research, focused on lying as an act of communication, has relied on assumptions about the audience to explain people's selection of deceptive messages (e.g. Ransom et al., 2019; Oey et al., 2019). Under an extreme formulation, these two accounts are fundamentally inconsistent: under the individual-based account, people are deeply inward-looking, considering only themselves, and their own values when choosing whether and how to lie; under the audience-based account, on the other hand, people consider the listener when making the same choices.

The current work pitted the individual- and audience-based accounts in direct competition, by considering how people lie when there is an explicit mismatch between their own prior beliefs and their estimates of their audience's prior beliefs. We introduced a novel dyadic lying game, in which a partially observable world state guides the players to have asymmetric beliefs. In these settings, we found that peoples' lies are better predicted by beliefs about their audience, as opposed to beliefs about themselves.

This study focused on beliefs about the world, one component of reasoning that drives behavior. Beliefs about the world can be asymmetric, which functions well to contrast individual- and audience-based accounts. However, in real world settings, the strategic theory-of-mind speaker and the moral individually-focused speaker are not mutually exclusive. People may lie by primarily trading off maximizing their gain and avoiding audiences' detection, but they may secondarily avert downright unethical lies. Both cognitive mechanisms are likely weighed variably across contexts, just as the propensity to lie varies across experimental paradigms and laboratory versus field studies (Gerlach, Teodorescu, & Hertwig, 2019). Our results indicate that human lying behavior is not driven solely by individual-based considerations—people do take the audience into account when designing lies. However, that does not mean that there is no role for individual preferences—at the very least there is likely to be individual variation in aversion to lying, even though lies, when told, are strategically designed for the audience. Future work may more directly compare how other individual-based factors, like moral reasoning, trades off with audience-based factors.

Ultimately, this work has implications for how effective individual-based interventions for decreasing dishonesty (e.g. honesty pledges; Kristal et al., 2020) may be expected to be, compared to audience-based interventions (e.g. raising believed probability of detection). In addition, this work warns of people's potentially dangerous capabilities at exploiting others' idiosyncratic beliefs. For instance, fake news that seems jarringly false to some readers may have been effectively pitched to the beliefs of a select group of readers, e.g. alt-right news sources may design clickbait for their alt-right audience, not constrained by keeping stories plausible to a nonpartisan audience. With more information about the prior beliefs of the target audience, fake news may be more effectively targeted, rendering the audience more susceptible to its deception.

Overall, these findings show that people can spontaneously tune their lies to their estimates of their audiences' prior beliefs. These results support the claim that people may lie by primarily capitalizing on theory-of-mind to evade detection and exploit their audience's expectations.

Acknowledgments

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

References

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

Apperly, I. A., Riggs, K. J., Simpson, A., Chiavarino, C., & Samson, D. (2006). Is belief reasoning automatic? *Psychological Science*, *17*(10), 841–844.

Becker, G. S. (1968). Crime and punishment: An economic approach. In *The economic dimensions of crime* (pp. 13–68). London: Palgrave Macmillan.

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

Capraro, V., Schulz, J., & Rand, D. G. (2019). Time pressure and honesty in a deception game. *Journal of Behavioral and Experimental Economics*, *79*, 93–99.

Ding, X. P., Wellman, H. M., Wang, Y., Fu, G., & Lee, K. (2015). Theory-of-Mind training causes honest young children to lie. *Psychological Science*, *26*(11), 1812–1821.

Erat, S., & Gneezy, U. (2012). White lies. *Management Science*, *58*(4), 723–733.

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.

Gino, F., Ayal, S., & Ariely, D. (2009). Contagion and differentiation in unethical behavior: The effect of one bad apple on the barrel. *Psychological Science*, *20*(3), 393–398.

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

Jacobsen, C., Fosgaard, T. R., & Pascual-Ezama, D. (2018). Why do we lie? A practical guide to the dishonesty literature. *Journal of Economic Surveys*, *32*(2), 357–387.

Kristal, A. S., Whillans, A. V., Bazerman, M. H., Gino, F., Shu, L. L., Nina, M., & Ariely, D. (2020). Signing at the beginning versus at the end does not decrease dishonesty. *Proceedings of the National Academy of Sciences*, *117*(13), 7103–7107.

Lee, J. J., & Pinker, S. (2010). Rationales for indirect speech: The theory of the strategic speaker. *Psychological Review*, 785–807.

Levine, T. R. (2014). Truth-Default Theory (TDT): A theory of human deception and deception detection. *Journal of Language and Social Psychology*, *33*(4), 378–392.

Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, *45*(6), 633–644.

Montague, R., Navarro, D. J., Perfors, A., Warner, R., & Shafto, P. (2011). To catch a liar: The effects of truthful and deceptive testimony on inferential learning..

Oey, L. A., Schachner, A., & Vul, E. (2019). Designing good deception: Recursive theory of mind in lying and lie detection. In A. K. Goel, C. M. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st annual meeting of the cognitive science society* (pp. 897–903).

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

Shalvi, S., Eldar, O., & Bereby-Meyer, Y. (2012). Honesty requires time (and lack of justification). *Psychological Science*, *23*(10), 1264–1270.

Suchotzki, K., Verschuere, B., Van Bockstaele, B., Ben-Shakhar, G., & Crombez, G. (2017). Lying takes time: A meta-analysis on reaction time measures of deception. *Psychological Bulletin*, *143*(4), 428–453.

Verschuere, B., Meijer, E. H., Ariane, J., Hoogsteyn, K., Orthey, R., McCarthy, R. J., … Yildiz, E. (2018). Registered replication report on Mazar, Amir, and Ariely (2008). *Advances in Methods and Practices in Psychological Science*, *1*(3), 299–317.

Vrij, A., Fisher, R., Mann, S., & Leal, S. (2006). Detecting deception by manipulating cognitive load. *Trends in Cognitive Sciences*, *10*(4), 141–142.

Vrij, A., Granhag, P. A., & Porter, S. (2010). Pitfalls and opportunities in nonverbal and verbal lie detection. *Psychological Science in the Public Interest*, *11*(3), 89–121.