Designing good deception: Recursive theory of mind in lying and lie detection

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

Abstract

The human ability to deceive others and detect deception has long been tied to theory of mind. We make a stronger argument: in order to be adept liars – to balance gain (i.e. maximizing their own reward) and plausibility (i.e. maintaining a realistic lie) – humans calibrate their lies under the assumption that their partner is a rational, utility-maximizing agent. We develop an adversarial recursive Bayesian model that aims to formalize the behaviors of liars and lie detectors. We compare this model to (1) a model that does not perform theory of mind computations and (2) a model that has perfect knowledge of the opponent's behavior. To test these models, we introduce a novel dyadic, stochastic game, allowing for quantitative measures of lies and lie detection. In a second experiment, we vary the ground truth probability. We find that our rational models qualitatively predict human lying and lie detecting behavior better than the non-rational model. Our findings suggest that humans control for the extremeness of their lies in a manner reflective of rational social inference. These findings provide a new paradigm and formal framework for nuanced quantitative analysis of the role of rationality and theory of mind in lying and lie detecting behavior.

Keywords: deception; Theory of Mind; Bayesian reasoning; non-cooperative games; computational modeling

Introduction

The frank truth is that humans lie frequently, and the abilities to lie and detect lies are practical, but cognitively demanding, tools we develop over time (Vrij, Fisher, Mann, & Leal, 2006). Although much of the research on lying focuses on physical cues that give away lying (like facial expressions), both liars and lie detectors must consider not only the execution of lies (e.g. Vrij, Granhag, & Porter, 2010; Ekman, Friesen, & O'Sullivan, 1988) but also the informational content of lies. In our current era of endemic fake news (Allcot & Gentzkow, 2017), it is ever more critical that we develop an understanding of what cognitive processes contribute to deception and its detection.

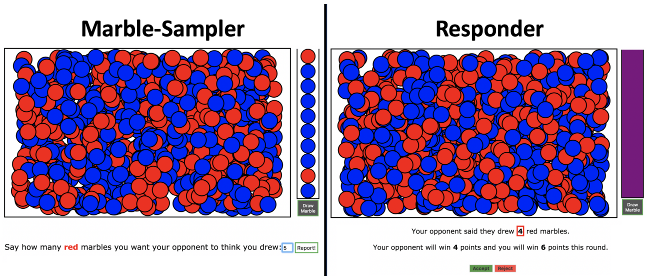
Lying *at all* requires believing that the recipient could have a belief different from your own, and thus lying has long been tied to theory of mind (ToM), or the understanding of others' mental states, such as beliefs. Children struggle with the ability to represent false beliefs and second-order beliefs conditioned on false beliefs (Wimmer & Perner, 1983; Talwar, Gordon, & Lee, 2007). This poor ToM in children should also make them terrible liars. Indeed, improvement in children's detection and production of lies appears to be directly related to the development of their ability to use ToM (Ding, Wellman, Wang, Fu, & Lee, 2015). To lie at all, we need to be able to entertain the possibility of a false belief in our interlocutor, however, successful deception requires a far more nuanced process of decision-making interacting with ToM inference.

We usually lie to benefit ourselves. For example, a male date-seeker may want to optimize his chances of attracting potential romantic interests by inflating his height on his online dating profile (Toma, Hancock, & Ellison, 2008). What height should he make up and report to accomplish this goal? A taller height might be more attractive in the eye of potential dates; so perhaps, he could choose the height of his favorite professional basketball player. However, being caught in a lie tends to be costly: he may jeopardize his trustworthiness. An overly tall height is likely to make his date more suspicious, so to decrease the chance of getting caught in a lie, he should not make the height too suspicious. How should he balance these competing pressures on his lie?

On the receiving end of a lie, it is advantageous for humans to be attuned to the detection of lies. Potential dates should want to detect the date-seeker's lie in order to discern whether he is a trustworthy human. But dates cannot haphazardly accuse others of lying, as a false accusation can also result in tarnishing the accuser's reputation. Both liars and lie detectors not only must navigate the constraints placed upon themselves, but they should also consider the other agent's perspective.

In the current study, we argue that good deception not only requires the use of ToM, but we make a stronger claim that good lie detectors evaluate, and good liars conjure, their lies under the assumption that their partner is a rational utility-maximizing agent. We formalize the role of rational and recursive social inference in the production and detection of deception. We argue that it is not only the ability to represent partners' false beliefs that distinguishes good liars from bad liars; rather, good liars balance maximizing reward with maintaining plausibility in their lies, such that liars can avoid having their lies detected by another agent. In order to maximize achievement of these goals, liars consider their partner's prior expectations, the likelihood of observations, and how these expectations shift in response to considering the other agent.

Traditionally psychological studies examining the role of ToM in deception are one-shot experiments. Examples of



Say how many **red** marbles you want your opponent to think you drew: 5

Your opponent said they drew **4** red marbles.

Your opponent will win **4** points and you will win **6** points this round.

Figure 1: Lying game. In both the marble-sampler and responder roles, participants see the distribution of marbles. (Left) Marble-samplers sample 10 marbles, then either lie or tell the truth about the number of red marbles sampled. (Right) Responders accept or call BS.

such studies are those in which children are instructed to not peek at a toy while the experimenter temporarily leaves the room, and children choose to either lie or tell the truth about peeking (Lewis, Stanger, & Sullivan, 1989; Talwar & Lee, 2002). Alternatively studies of dishonesty in the behavioral economics literature use quantitative measures but emphasize the tendency of people to cheat at an individual level, independent of how other agents affect their deception (e.g. Mazar, Amir, & Ariely, 2008). Taking inspiration from both designs, we developed a novel repeated dyadic stochastic game allowing us to focus on the quantitative, socially-motivated production and evaluation of lies.

Using Bayesian game-theoretic computational modeling and experimental methods, we argue that a well-calibrated, Bayesian ToM supports the production of believable lies and the detection of poorly-formed lies, and introduce a novel ideal observer model of deception.

Lying Game

To study how humans actually behave in lying situations, we developed a novel lying game that rewarded participants for strategic detection and production of lies (Figure 1).

In each round of the game, both players are presented with a box containing red and blue marbles, with some proportion p of red marbles. Players alternate between playing as the marble-sampler or the responder. The marble-sampler randomly samples 10 marbles from the box, of which k are red. However, the sampled marbles are occluded from the responder, so the responder cannot see the true distribution of sampled marbles. The marble-sampler chooses a number k^* to report as the number of red marbles they want their opponent to *think* they sampled. The marble-sampler could choose to (a) tell the truth and report the true number of red marbles sampled, or (b) lie and report a false number of red marbles sampled. The responder then has the opportunity to either (A) accept the reported value or (B) reject it as a lie (i.e. call BS).

Both the marble-sampler's decision to (a) tell the truth or (b) lie about the number of red marbles sampled, and the responder's decision to (A) accept or (B) reject the reported value impact each player's payoff (Table 1). If the reported number of red marbles sampled k^* is accepted, the marble-sampler receives k^* points and the responder received $10 - k^*$ points. If the responder rejects the reported number then the payoffs depend on whether or not it was a lie: if the reported number is the truth ($k = k^*$), the marble-sampler gets the k^* points, and the responder pays a penalty of $-k^*$; if the reported number is a lie ($k \neq k^*$) then the responder gains k^* points, while the marble-sampler pays a penalty of $-k^*$. Altogether, this game sets up a reward function that motivates marble-samplers to lie, but not be caught, and motivates the responder to call out egregious lies, but avoid false accusations.[1]

		Marble-Sampler	
		$k = k^*$	$k \neq k^*$
Responder	$BS = 0$	$2k^* - 10$ $10 - 2k^*$	$2k^* - 10$ $10 - 2k^*$
	$BS = 1$	$2k^*$ $-2k^*$	$-2k^*$ $2k^*$

Table 1: Players' payoff differential (player - opponent points). Utility is determined by reported k^* and whether BS was called. Values to the right of the diagonal in cells indicate points awarded to the marble-sampler, while values to the left are awarded to the responder.

Models

No-Theory-of-Mind Model

As a baseline, let's consider a model that has no model of the opponent, or believes that the opponent is effectively random. In deciding upon what number to report (k^*), such a model does not consider the behavior of the opponent, and would simply lie with probability $1 - p$. Moreover, when it lies it would either sample uniformly from values larger than the truth (k), or it would simply pick the largest value (10 – as this is expected-value maximizing response under the assumption that the opponent calls BS at random). This is the best that an agent that has no model of their opponent could do. This no-theory-of-mind model makes a qualitative prediction about lying behavior, such that the expected value of k^* increases linearly as a function of k.

Likewise, a lie-detector that has no model of their opponent, and thus believes them to be random, would only consider the probability of k^* under the true world distribution of $P(k)$. Since this model does not consider the motives and payoffs for their opponent, it would amount to playing the game without knowing the opponent's payoff structure, e.g. whether they would receive points for red or blue marbles. If the marble-sampler were to say that they sampled one red marble when $p = 0.5$, the responder may call BS, simply because such a value is unlikely to occur by chance. This lie detector amounts to conducting a two-tailed hypothesis test. It computes what is statistically significant under a binomial test and calls BS on all k^* that have a p-value $< \alpha$. Regardless

[1] Code available at `github.com/la-oey/Bullshitter`

of α, this lie-detector would call BS on all reports of unlikely k^*, and would thus have a U-shaped lie-detection profile.

Oracle Model

Alternatively, suppose we have a theory-of-mind model that has a *perfect* model of its opponent (i.e. it has an oracle-like omniscience over its opponent's probability to lie and detect lies). This model does not require recursive social inference as a simple first-order inference will suffice, given that they have already perfectly adapted their model of the opponent. It is critical to understand how such a model would behave, as this exemplifies an ideal agent.

To accomplish this, we developed an inferential model of deception, which we term the oracle model, whose opposing agent lies and detects lies using the algorithms from the AI that participants competed against in our lying game in both experiments.

When detecting lies, the AI computes $P_D(BS \mid k^*)$ using the cumulative binomial probability of k^*, $P(X \leq k^*) = \sum_{x=0}^{k^*} Binomial(x \mid p, 10)$ centered at 0.5 when $k^* = 5$. To compute $P_L(k^* \mid k)$, the AI randomly samples a potential k^*, $\hat{k}^*$, from a binomial distribution. If $\hat{k}^*$ is greater than the true k, it lies and uses $\hat{k}^*$ as its reported k^*. Otherwise, it tells the truth by using the true k as its reported k^*.

As participants in our behavioral experiment iteratively competed against this very same non-inferential algorithm over several trials, it seems viable that participants may become perfectly calibrated to the algorithm that their opponent operated upon. In that case, human behavior would rationally match the predictions of the oracle model performing inference over the AI.

Recursive Theory-of-Mind Model

Finally, we consider a model of an ideal observer who does not know *a priori* the behavior of their opponent, but can estimate it from first principles, on the assumption that their opponent is as rational as they are, and is also trying to anticipate their opponent's behavior. This amounts to paired, adversarial ideal observers in which liars L and lie detectors D act as competing rational utility-maximizers. Our model builds on previous Bayesian frameworks of social cognition and communication (Baker, Saxe, & Tenenbaum, 2009; Frank & Goodman, 2012). Both agents perform inference over one another, i.e. L determines what number to report based on his prediction of D's tendency to call BS for different reported numbers, and vice versa. Both agents assume the other agent is acting rationally, namely the other agent is performing optimally given their goal to maximize their own utility. Furthermore, this process of performing inference over the other agent's actions is recursive. In other words, L decides upon his action based on what he believes D will do in light of what she believes L will do, etc. As infinite recursion is memory delimited, our model implements a decay function that breaks the chain of recursion with some degree of probability and implements a base case.

L is constrained by two competing goals: (1) gain (i.e. the agent wants to gain the highest reward possible given that they successfully deceive their partner), and (2) believability (i.e. the probability of having a lie go undetected, which is constrained by the extremeness of their lie). Meanwhile, the competing goals of D include to (1) successfully detect lies, while (2) avoiding falsely accusing their partner of lying when they are in fact telling the truth. The adversarial nature of the agents' goals is captured in the inverse relationship of the utility values for both L and D in our lying game.

Given some true state of the world k, L asserts to D that the state of the world is k^*. If k^* is not equal to k, L is telling a lie, otherwise L is telling the truth. D then sees the reported k^*, and responds by choosing whether to challenge the veracity of k^* by calling $BS = 1$, or accepting k^* as stated ($BS = 0$).

Our formalization of deception is represented as a zero-sum game. We assume that the probability of an action follows a Luce choice rule based on the expected utility of the action relative to alternative actions, with softmax parameter α (Luce, 1959):

$$P(A) = \underset{A}{softmax}(\mathrm{EV}[A]) = \frac{\exp(\alpha \mathrm{EV}[A])}{\sum_{A'} \exp(\alpha \mathrm{EV}[A'])} \tag{1}$$

D chooses to call BS following a Luce choice rule weighting of the expected value of the two options: calling BS, or accepting k^*:

$$P_D(BS \mid k^*) = \underset{BS}{softmax}(\mathrm{EV}_D[BS \mid k^*]) \tag{2}$$

The expected value of calling BS is obtained by marginalizing over the possibilities that $k^* = k$ (here abbreviated as $T = 1$), and $k^* \neq k$ ($T = 0$):

$$\mathrm{EV}_D[BS \mid k^*] = \sum_T \mathrm{u}_D(BS; k^*, T) P(T \mid k^*) \tag{3}$$

where $\mathrm{u}_D(BS; k^*, T)$ is the payoff for D associated with a particular BS response, given k^* and whether or not it corresponds to the true k (T).

The probability of a given k^* being true is given by

$$P(T \mid k^*) = P(k^* = k \mid k^*) = \frac{\sum_k P(k) P_L(k^* \mid k) P(k = k^* \mid k, k^*)}{\sum_k P(k) P_L(k^* \mid k)} \tag{4}$$

relying on the prior probability of k (here: $P(k) = Binomial(k \mid p, 10)$), and the probability that L would produce a given k^* in response to seeing a particular k, $P_L(k^* \mid k)$. Thus, calculating the expected value of calling BS, and choosing whether or not to call the lie requires an estimate of how L is likely to behave.

L, in turn chooses k^* based on a softmax weighting of the expected value of different responses,

$$P_L(k^* \mid k) = \underset{k^*}{softmax}(\mathrm{EV}_L[k^* \mid k]) \tag{5}$$

with the expected values given by:

$$\mathrm{EV}_L[k^* \mid k] = \sum_{BS} \mathrm{u}_L(k^* \mid BS, k^* = k) P_D(BS \mid k^*) \tag{6}$$

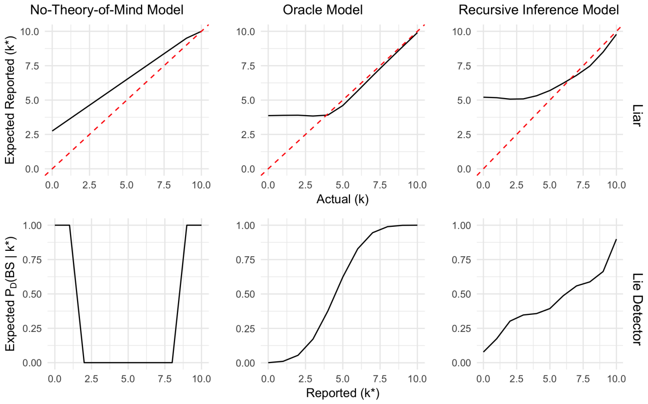


Figure 2: Computed predictions across all models (no-theory-of-mind, oracle, and recursive inference) when $p = 0.5$. Top row displays the liar's predicted performance: reported value as a function of the true value. The red dashed line indicates reported values that would be true. Bottom row displays the lie detector's predicted performance: conditional probability of calling BS by the reported value k^*.

where $\mathrm{u}_L(k^* \mid BS, k^* = k)$ is the payoff for the L when reporting k^* given that BS was called, and whether that k^* was a lie. Calculating these expected values requires that L consider $P_D(BS \mid k^*)$ – the probability that D would call BS for a particular reported k^*.

Thus the expected values of various choices for L depends on his beliefs about D, and the expected values of calling BS for D, depends on her beliefs about L. This would yield infinite recursion, so in practice we assume that L's model of D has some probability λ of simply returning a constant $P_D(BS \mid k^*) = c$.

Model Predictions

In Figure 2, we computed predictions from each of the three models about the performance of liars' reported k^* given true k and lie detectors' $P(BS \mid k^*)$ given the reported k^*.

The oracle and recursive inference models make qualitatively similar predictions. For lying, above a certain k, reported values tend to fall on the identity line, indicating that beyond that value, lying is imprudent. For values of k below the average, it is better to lie with a false report of an average outcome (here, $\mathrm{E}[k^*] = 5$ for small values of k). This threshold value in the oracle model is lower than the one in the recursive inference model. This pattern of lying seems to reflect the liar's attempt to balance the gain of lies and the risk of detection from reporting improbable values. For lie detecting, both models predict a sigmoidal pattern, calling BS more often as k^* increases. The fact that both the oracle and recursive inference models appear similar along both the liar and lie detector behaviors indicates that the recursive inference model can emulate the same behavior as the oracle model, despite having no information about the specific behavioral policies of the opposing agent.

In contrast to the theory-of-mind (oracle and recursive inference) models, the no-theory-of-mind model only reduces lying on account of a ceiling effect; thus making k^* a linear function of k. As a lie detector, the no-theory-of-mind model does not consider the reward function of the liar, and thus predicts that both extremely high and extremely low reported values would be called out as lies.

We qualitatively tested these predictions from the theory-of-mind and non-theory-of-mind models in experiment 1. In experiment 2, we tested how manipulating the prior probability of sampling k by varying p would influence human lying and lie detecting behavior. Under the assumption that liars and lie detectors behave rationally, we would expect to see that their behavior would be robust to changes in the probability of the world.

Experiment 1

Participants

We recruited 193 UC San Diego undergraduate students to participate in an online study for course credit.

Procedure

There were a total of 40 trials, with the player acting as the marble-sampler in the initial trial, and then switching roles between each trial, resulting in 20 trials as the marble-sampler and 20 trials as the responder. Participants were instructed to "beat [their] opponent into the ground by winning by the highest point differential possible," in order to motivate participants to successfully lie and detect lies throughout the task. The distribution of marbles was uniform, such that there were 50% red and 50% blue marbles ($p = 0.5$)

Results

When in the marble sampler role, participants showed a non-linear pattern of drift from the truth with lower k values, as shown in the top of Figure 3. We find that this pattern of lying in a positive utility direction for the liar, i.e. above the red line, occurs at lower numbers up until the actual marbles sampled is equal to 5 (i.e. the expected mean).

When in the role of responder, participants' results showed a sigmoidal trend, as shown in the bottom of Figure 3. Both the liar and lie detector pattern of results provide evidence against the no-theory-of-mind model and instead support the oracle and recursive inference models.

It should be noted that due to the nature of binomial distributions, sampling a low number (or high number) of red marbles is rare. As the AI was set up in such a way that the computer tends to lie toward the mean value when the sampled number of marbles is low, this produced a low probability of the computer reporting a low number of red marbles sampled. As a result, there were only a small number of data points available to determine how people detect lies under those conditions. To help offset the wide variance resulting from low counts across k^*, we converted counts to proportion using $(n_{BS=1} + 1)/(n + 2)$ and for all figures, we included points in which there were greater than three observations for a given value along the x-axis.

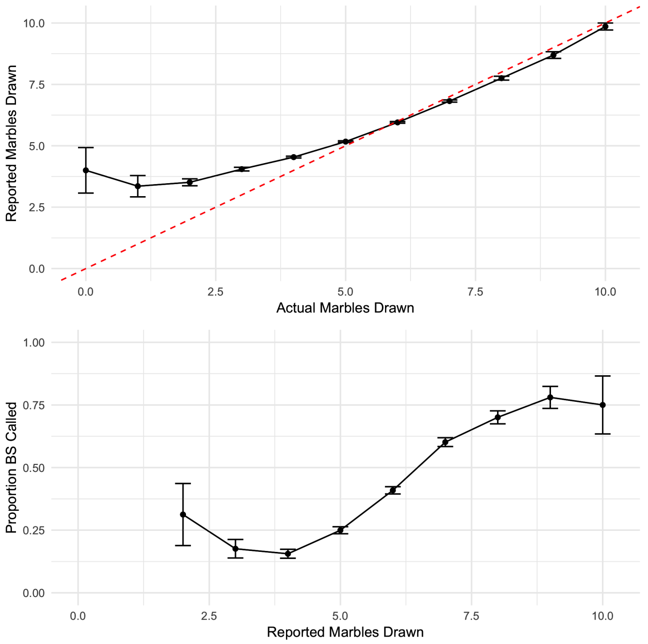


Figure 3: Results from experiment 1: (Top) Marble-sampler's reported number versus actual red marbles sampled. (Bottom) Responder's proportion of calling BS by the reported number of red marbles sampled.

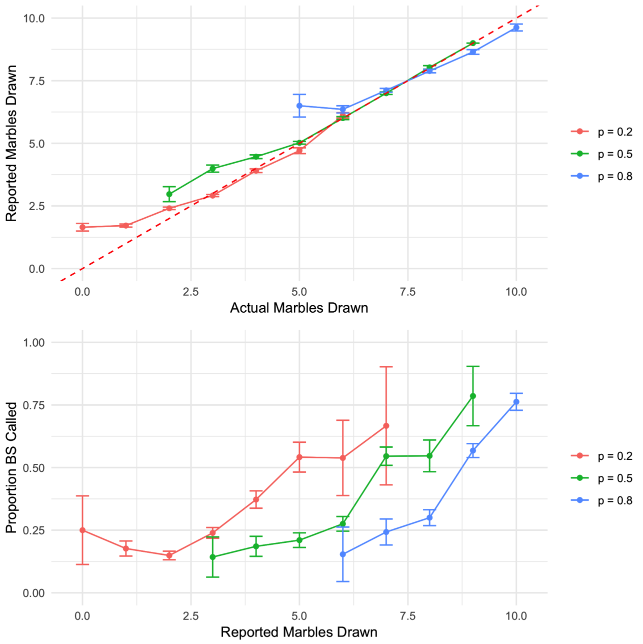


Figure 4: Results from experiment 2: (Top) Marble-sampler's reported versus actual number of red marbles sampled, by changes in the distribution of red-to-blue marbles (indicated by color). (Bottom) Responder's proportion of calling BS by the reported number of marbles sampled.

Experiment 2

We predicted that if people were making flexible, rational inferences, they would be able to flexibly take into account the distribution of marbles in the population, both in the lies they generated and in their detection of others' lies. In lying, we expect a shift in the point on k at which reported values drift from the truth. Similarly, in lie detecting, we expect a change in the tolerance for different k^* values, resulting in a horizontal shift of the BS calling function.

Experiment 2 used a similar design to experiment 1, except that crucially we manipulated the prior probability of k and removed feedback about the other agent's actions. By eliminating feedback, we hoped to distinguish between whether participants are simply adapting to the strategy of the other agent from this feedback, or if participants are performing inference on the other agent's decision process.

Participants

We recruited 86 UCSD undergraduates. Fifteen participants failed to meet the attention check criteria, which entailed accurately answering greater than 75% of the 12 comprehension questions disbursed throughout the experiment. This left 71 participants in our final pool.

Procedure

The procedure for experiment 2 was similar to experiment 1, except we varied between-subject the probability distribution from which the marbles were sampled. There were three conditions: the (red-to-blue) distribution of marbles was either 50-50 ($n = 20$), 20-80 ($n = 32$), or 80-20 ($n = 19$). Participants were not explicitly told about the distribution of marbles; rather they gathered the value of $P(k)$ via visual observation (i.e. the distribution of marbles in the box). In addition, the number of trials increased to 80 trials.

Participants also received no feedback about player decisions between each trial. Thus, the marble-sampler no longer received feedback about whether the responder called BS, and the responder no longer received about whether the marble-sampler lied or not. To ensure that participants understood the payoff structure, participants completed four pre-task practice trials with feedback, i.e. players' decisions, points earned a given trial, and cumulative score, after each trial. These practice trials were included to demonstrate to the participants that the other agent was generating lies or evaluating the participants' lies, and to establish the game's payoff matrix. After the practice trials, participants were told they would no longer receive feedback after each trial, instead only seeing the cumulative score every fifth trial.

Lastly, we used a new payoff structure, in order to contend with a puzzling characteristic of the payoff structure used in experiment 1. In particular in experiment 1, when $k^* \leq 2$, the responder received a lower relative payoff for successfully calling BS on a k^* lie than accepting k^*. Therefore, it would be in the responder's best interest to accept k^* even if they were to believe the marble-sampler was lying. In the $p = 0.2$ condition, the expected value of $k^* = 2$, suggesting that this condition may be affected by the unusual payoff at lower k^* values. To contend with this issue, experiment 2's new payoff structure resulted in a relative gain of 10 for the responder (and a relative loss of 10 for the marble-sampler) whenever the responder successfully called BS on a lie. Meanwhile, falsely accusing the marble-sampler of lying resulted in a -5 penalty (deducted from the points they would have received had they accepted) for the responder.

Results

Overall we found that participants calibrate their lies, as well as their lie detection, based on the probability structure of the world. Firstly, when examining the lies given by the marble-sampler, we can see a drift in the reported k^* across all conditions, following the pattern of response seen in experiment 1. This point of drift shifts in accordance with the condition, such that the shift in condition $p = 0.2, 0.5, 0.8$ occurs around $k = 3, 5, 7$, respectively.

Secondly, in examining lie detecting behavior, we found that participants shift their judgments of which k^* values they called out as a lie, based on the probability distribution of marbles in the population. We used a mixed-effects logistic regression model to describe the probability that BS is called as a function of the reported number of red marbles sampled k^* and the marble distribution p dummy coded with $p = 0.5$ as the reference group. We found a significant main effect of both reported number of red marbles sampled ($\hat{\beta} = 0.723$, $z = 9.032$, $p < 0.0001$) and marble distribution in the $p = 0.2$ condition ($\hat{\beta} = 2.069$, $z = 3.081$, $p = 0.002$) and in the $p = 0.8$ condition ($\hat{\beta} = -5.823$, $z = -4,714$, $p < 0.0001$). These results not only suggest that people's detection of lies varies as a function of the reported value, but people also calibrate their BS calling depending on the probability of the world.

Using the estimated β values, we computed the number of marbles k^* at which lie detectors would call BS 50% of the time ($P_D(BS \mid k^*) = 0.5$). These thresholds varied systematically across the different marble distributions ($p = 0.2$: 5.236; $p = 0.5$: 7.327; $p = 0.8$: 8.762). The decision boundary shifts to higher k^* values as p increases. This result suggests that lie detectors change their BS calling behavior as a function of their prior expectations about the distribution of the world.

Discussion

In this paper, we report evidence that people lie, and detect lies, in ways that are well-captured by an adversarial recursive Bayesian model. We argue that good liars not only require an ability to represent the idea that others might have mental states different from their own, but they make inferences about the beliefs and actions of their interlocutor to successfully evade detection. In determining what utterances to call out as lies, good lie detectors must rationally consider the goals and utilities of their interlocutor and statistical information about the probability structure of the world.

We introduced the oracle model, in which the model has perfect information about how its opponent behaves. We compared the oracle model to an ideal observer model that does not know the opponent's exact behavioral policies – as is the case in real-world lying – but must instead deduce the opponent's behavior from first principles. This ideal observer assumes that the opponent is rational, and thus, estimates the opponent's behavioral policies by performing recursive social inference. We found that both the oracle and recursive inferential models make qualitatively similar predictions about both lying (i.e. non-linear lies as a function of the true value) and lie detecting (i.e. logistic pattern of calling BS as a function of the lie). This lack of distinguishing predictions across these two models suggests that even though the recursive inferential model lacks the omniscience of the oracle model, it can reproduce qualitatively the same behavior with a far sparser explicit representation of the other agent.

The oracle and recursive theory of mind models are contrasted with an agent that has *no model* of the opposing agent. This agent lies by only considering rewards (and not the opponents' reaction), and detect lies by only considering what is improbable (and not what lies would favor the opponent). This agent makes qualitatively different predictions about both lying and lie detecting behavior. We then tested these model predictions by examining human behavior in a novel lying game. The empirical results suggest that the recursive inferential model of deception capture how human liars choose which lie to tell: they tend to choose lies that are not too implausible. Likewise, we find that lie detecting behavior is consistent with recursive ToM and is calibrated to the probability of the sample under the prior distribution of the world.

To better determine how recursion in these rational models maps onto human behavior, one natural future direction would be to have participants compete against each other in this game. Is it truly the case that liars assume lie detectors are reasoning rationally about the liar, and lie detectors assume liars are reasoning rationally about the lie detector?

In the current experiments, people played against a computer opponent with fixed, non-adaptive behavior. Perhaps over the iterative trials, participants perfectly adapted their model of the other agent, such that they knew how it would behave as a liar and lie detector–essentially acting like the oracle model, with no need for any more than first-order ToM, or ToM over an agent who does not assume rationality about the other agent. Do participants typically perform recursion, or do people only perform first-order ToM? Our first pass at providing evidence against this alternative hypothesis is shown in our second experiment in which the lack of feedback about the opponent's behavior requires players to generate a model of the agent without actual knowledge about the agent's behaviors. Given this lack of feedback, it would be far more difficult to develop an accurate non-inferential generative model of the other agent.

In summary, in the study we present here, we propose and contribute empirical evidence that liars and lie detectors act as rational utility-maximizing agents. Liars and lie detectors choose how to lie and when to call out lies under the assumption that the other agent is also behaving rationally. These findings provide a stepping stone for novel quantitative approaches to studying deception.

Acknowledgments

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1650112 to LAO, and National Science Foundation Grant No. BCS-1749551 to AS.

References

Allcot, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–236.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*, 329–349.

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.

Ekman, P., Friesen, W. V., & O'Sullivan, M. (1988). Smiles when lying. *Journal of Personality and Social Psychology*, *54*(3), 414–420.

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

Lewis, M., Stanger, C., & Sullivan, M. W. (1989). Deception in 3-year-olds. *Developmental Psychology*, *25*(3), 439–443.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York: Wiley.

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.

Talwar, V., Gordon, H. M., & Lee, K. (2007). Lying in the elementary school years: Verbal deception and its relation to second-order belief understanding. *Developmental Psychology*, *43*(3), 804–810.

Talwar, V., & Lee, K. (2002). Development of lying to conceal a transgression: Children's control of expressive behaviour during verbal deception. *International Journal of Behavioral Development*, *26*(5), 436–446.

Toma, C. L., Hancock, J. T., & Ellison, N. B. (2008). Separating fact from fiction: An examination of deceptive self-presentation in online dating profiles. *Personality and Social Psychology Bulletin*, *34*(8), 1023–1036.

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.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*, 103–128.