Reasoning about knowledge in lie production

Tan Zhi Yi[1], **Julian Jara-Ettinger**[2], **& Marlene D. Berke**[2]
e0565013@u.nus.edu, julian.jara-ettinger@yale.edu, marlene.berke@yale.edu
[1]Yale-NUS College
[2]Department of Psychology, Yale University

Abstract

Theory of Mind enables us to represent and reason about other people's mental states like beliefs and knowledge. By considering what other people know, this allows us to strategically construct believable lies. Previous work has shown that people construct lies to be consistent with others' beliefs even when those beliefs differ from their own. However, in most real world cases, we don't know everything that the other person knows. We propose that to produce believable lies, the sender considers what private information the receiver may have. Here, we develop our theory into a computational model and test it in a novel paradigm that allows us to distinguish between knowledge shared between the lie sender and receiver and knowledge private to the receiver. Our main model successfully captures how people lie in this paradigm over alternative models. Overall, our work furthers our understanding of human social cognition in adversarial situations.

Keywords: Theory of Mind; Lying; Deception; Knowledge; Social Cognition

Introduction

Imagine the following scenario: your partner asked you to buy the kind of soda that they like, and reminded you several times, but you still forgot. You really do not want to admit that you were selfish and forgetful enough to only buy the kind that you like, so you start to think of excuses. You could say it was out of stock—but would they believe you? What if they've been to the store recently? Maybe you could say that the store was closed—but what if they know the store's hours? In many cases like this one, where we might want to lie, we don't know exactly what our audience knows. So what do we do? Do we simply ignore this problem? Or do we still try to reason about what they may or may not know?

Theory of Mind is the ability to represent and reason about other people's mental states like their beliefs (Gopnik, Meltzoff, & Bryant, 1997; Premack & Woodruff, 1978; Wellman, 2014), and it is central to the social skill of strategic lying (Oey, Schachner, & Vul, 2023). The abilities to reason about false beliefs and to actively deceive others are unique to the human species (Gouzoules & Gouzoules, 2002; Horschler*, Berke*, Santos, & Jara-Ettinger, 2023; Karg, Schmelz, Call, & Tomasello, 2015; Martin & Santos, 2016; Schmelz & Call, 2016; c.f., Krupenye & Call, 2019; Hall & Brosnan, 2017). Furthermore, lying emerges simultaneously in children as Theory of Mind (Lavoie, Leduc, Arruda, Crossman, & Talwar, 2017; Reddy, 2007), and training children's Theory of Mind has even been shown to increase their lying (Ding, Wellman, Wang, Fu, & Lee, 2015). And it makes sense that strategic lying involves ToM, as lying is about manipulating mental states.

One of the central challenges to telling strategic lies is making them believable to the audience. This difficulty is part of why strategic lying is so cognitively demanding—much more so than telling the truth (Debey, De Houwer, & Verschuere, 2014; Van't Veer, Stel, & van Beest, 2014; Verschuere, Köbis, Bereby-Meyer, Rand, & Shalvi, 2018). Recent work showed that people design lies so as to be consistent with what they know to be in the receiver's beliefs, even when the receiver's beliefs differ from their own (Oey et al., 2023). However, in most everyday scenarios, we don't know all of what our audience knows. In real life, the receiver's knowledge is rarely a subset of our own, but rather, a set of beliefs that shares some overlap with our own (shared knowledge) and some disjoint (private knowledge). Here, we investigate whether people actually consider the receiver's possible private knowledge when constructing strategic lies, or if considering just shared knowledge is sufficient to explain when and how people lie.

In this paper, we formalize these theories as computational models and pit them against each other, testing their ability to explain human behavior both at the group and individual levels. We present a novel paradigm called *Bears and Tracks*, which allows us to distinguish between knowledge that's shared between the lie sender and receiver, the knowledge that's private to the sender, and the knowledge that's private to the receiver, in a way that previous paradigms (e.g., Oey et al., 2023) have not been able to capture. This allows us to test new questions about the cases in which we really do represent others' private knowledge vs. simply use our shared knowledge to produce believable, strategic lies.

Paradigm

The purpose of our novel *Bears and Tracks* paradigm is to test lie production as the contents of knowledge shared between the sender and receiver and knowledge private to each vary separately. Our cover story goes as follows: an oil company wants to build a pipeline through a national park. However, a protected species of bears may live in the area. Participants play the role of the oil company, and they are incentivized to under-report the number of bear dens that they find so as to be allowed to develop the land. A surveyor from the Environmental Protection Agency (EPA) surveys one third of the

land by themself, and one third of the land together with the oil company. The oil company surveys the remaining third by itself, and then has to report to the surveyor the number of bear dens that they found in the third of the land that they surveyed alone. The EPA surveyor receives the oil company's report and decides whether to accept the report or, if they suspect that it is a lie, to investigate it. Investigations always lead to the truth—determining how many dens were in the section the company surveyed, and whether or not the company lied. Participants are told that they will win if their company is allowed to develop the land, which can only happen if two or fewer bear dens are reported in the entire area and if they are not caught lying. If they are caught in a lie or if the overall report has two or more bear dens, then they lose.

What makes the paradigm interesting is that the bears sometimes leave tracks, and the tracks can be anywhere in the park, not just in the third where the den is. This means that, in a situation like the one shown in Fig. 1, the participant should be careful about lying by reporting zero bear dens, for fear the receiver will call out the lie based on the track in shared knowledge.

Further consider the situation in Fig. 1. The bear track is in the shared knowledge of the sender and the receiver, while the bear den is known only to the sender. The sender doesn't know what else, besides the track, the receiver might know. What should the sender report? If the sender chooses to lie and report zero bear dens, they run the risk that, if there are no bear dens in the section that only the receiver can see, their lie will be caught. For that reason, it might be prudent to tell the truth. However, if the participant is more concerned about the possibility of a bear den in the receiver's section, then it may be worth the risk to lie. However, there is absolutely no conceivable reason to lie by reporting two bear dens. This is the kind of reasoning that our model aims to capture.

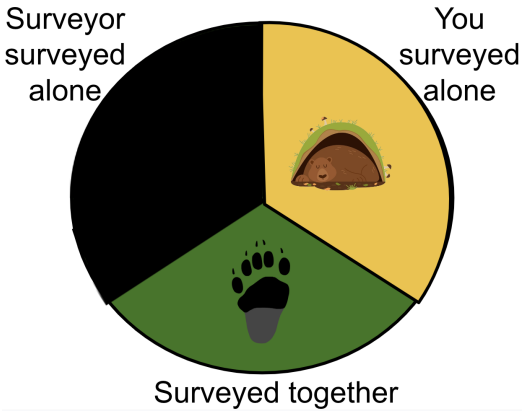

Figure 1: An example diagram of a scenario from the experiment. The blacked out third depicts the section that the EPA surveyor surveyed alone. The participant does not get to observe what the EPA surveyor found, so this section is always empty. The green third shows the area that the EPA surveyor and the oil company survey together, which currently has one bear track. The yellow third shows the area that the oil company surveys by itself, which currently has one bear den. Participants are tasked with reporting the number of bear dens in the sections that they surveyed by themselves (yellow), and deciding whether and how to lie.

Computational Model

Generative model of lie construction

Our model represents beliefs as a probability distribution over possible world states, and mental state inference as Bayesian inference over a generative model, both of which are standard (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017). In the context of our paradigm, there are no observable actions from which to infer the surveyor's beliefs, but a distribution over their possible beliefs can still be inferred based on the part of the world observable to participants and the causal structure of how world states are produced in this task.

In this model, the sender chooses how many bear dens to report based on estimating the expected value of each possible report that they could give. The expected value of a false report has two main components: its believability (which determines the probability that the EPA will investigate), and the reward if the report is accepted. The expected value of a true report is simply its reward, since it does not matter for the sender whether or not the EPA investigates a true report.

The believability of the report depends on the receiver's beliefs. Let the random variable S_i denote the contents of section i—the bear dens X_i and tracks T_i, so $S_i = \{X_i, T_i\}$. The receiver knows the contents of the section that they survey by themself, $S_1 = s_1$, and the contents of the section that they survey together with the sender, $S_2 = s_2$. The believability of the proposition $X_3 = prop$ is simply its conditional probability given what the receiver knows: $P(X_3 = prop | S_1 = s_1, S_2 = s_2)$. The EPA's probability of *not* investigating is assumed to equal the believability.[1]

The reward of a report if accepted is simply an indicator for whether the win condition of the aggregated report containing fewer than two dens is met: $\mathbb{1}(x_1 + x_2 + prop < 2)$. For a false report, its expected value is its believability times its reward, and for a true report, its expected value is its reward. Given the contents of S_1, the expected reward of reporting a proposition *prop* is:

$$EV(prop|S_1 = s_1) = \begin{cases} P(X_3 = prop \,|\, S_1 = s_1, & \text{if} \\ \quad S_2 = s_2) \times & X_3 \neq prop \\ \quad \mathbb{1}(x_1 + x_2 + prop < 2), & \\ \mathbb{1}(x_1 + x_2 + prop < 2), & \text{otherwise} \end{cases}$$

[1] This is an assumption—the reward structure of the task for the receiver is left unspecified, but different reward structures could result in different criteria for the receiver rejecting a report and choosing to investigate. This could be accommodated in this model by applying some transformation to the conditional probability, like $+0.1$ if the receiver is a bit more conservative about rejecting a report (perhaps because investigation is costly), or -0.4 if the receiver is extremely liberal and willing to investigate.

However, the sender does not know the contents of the receiver's private knowledge S_1, and therefore has to infer it given what they do know. The sender infers the distribution $P(S_1|S_2 = s_2, S_3 = s_3)$ by Bayesian inference, which is approximated in our model via Markov Chain Monte Carlo using Metropolis Hastings. Using the inferred distribution over S_1, the sender calculates the expected value for each possible report by marginalizing over S_1: $EV(prop) = \sum_{s_1} EV(prop|S_1 = s_1)P(S_1 = s_1|S_2 = s_2, S_3 = s_3)$.

The model predicts that participants will choose what to report according to a softmax over the expected values, parametrized by the inverse temperature parameter τ (Luce, 1959). This captures the assumption that participants will rationally select reports according to their expected values.

Priors over world states

The bear dens are assumed to be independently and randomly distributed over the park, thus following a Poisson distribution with rate parameter λ_{bears}, and truncated with an upper limit of two dens per third. (This upper limit helps keep the state space constrained, so computation is tractable.) The distribution of bear tracks depends on the number of bear dens in the whole park, but each track is otherwise independent of the others. Therefore, the tracks follow a Poisson distribution where the rate depends on the number of bear dens in the park: $n_{dens} * \lambda_{tracks}$. This distribution was also truncated with an upper limit of two tracks per third.

Alternative models

Ignoring the receiver's private knowledge It is possible that people might represent the knowledge that they share with the receiver, but not the knowledge that the receiver has that they lack. This would be like using a simplified Theory of Mind model, perhaps to reduce computational and cognitive costs. To test this theory, we lesioned S_1 out of the main model. This eliminated the need for Bayesian inference over S_1, and thus simplified the computations. We call this model Shared Knowledge Only as, according to this model, the sender represents the receiver as only having access to the knowledge that the two of them share.

Cost to lying So far, the models described above rationally (but probablistically) give the report that is expected to be most beneficial. In cases where it is impossible to win and the expected value of each report is 0, the models would give a random report. However, intuitively, this doesn't quite seem right. Why would we lie when there is nothing to gain? An agent that engages in frivolous, unnecessary lying is one that we might call a pathological liar.

To address this, we introduce a cost to lying. Reports other than the truth carry with them a cost, which impacts their expected values. When this parameter is absent (or equivalently, set to 0), we call this a Pathological Liar model. The main model with $cost_{lying} = 0$ is named Main x Pathological Liar, and the Shared Knowledge Only model with $cost_{lying} = 0$ is named Shared Knowledge Only x Pathological Liar.

Fitting parameters

The models described so far have four parameters: λ_{bears}, λ_{tracks}, τ, and $cost_{lying}$. In order for the models to quantitatively capture participants' behavior, the values of the λs have to align with participants' priors over the prevalence of bear dens and tracks, the value of τ has to capture the randomness in participants' choices, and the value of $cost_{lying}$ has to capture participants' aversion to lying. These four parameters were fitted separately for each model so as to maximize its Pearson correlation with participant responses.

Behavioral Experiment

To differentiate between models and therefore learn about which theories of lie production best predict participant behavior, we ran an experiment where participants responded to dozens of scenarios drawn from our *Bears and Tracks* paradigm. The preregistration and all materials can be found on our OSF repo: `https://osf.io/qaxtg/?view_only=ccff1e6603b3484d81c4c385fb4ca9c0`

Participants

100 U.S. participants (Age: mean = 39.6 years, range = 19-72 years) were included in our analysis. In total, 107 U.S. participants were recruited on Prolific, but seven were excluded from our analysis for reporting paying $< 90\%$ attention.

Stimuli

Stimuli consisted of 32 diagrams depicting scenarios varying the number of bear dens and tracks in the partition observable to both participants and the receiver, and in the partition observable solely to the participant. Since both tracks and bear dens had an upper limit of two per partition, there were nine different combinations of numbers of tracks and bears that could be in one partition (zero bears and zero tracks, one bear and zero tracks, etc), yielding $9^2 = 81$ possible different scenarios that the participant could observe. Because the task instructions were ambiguous about whether the receiver believed that the number of bear tracks could possibly outnumber the bear dens, we filtered out the scenarios where there were two bear tracks in common ground, leaving 54 scenarios. Of these 54 scenarios, 18 were scenarios where winning was trivially impossible (with two bear dens in shared knowledge). To test how people would behave (and whether they would lie) when winning was impossible, we included six of these scenarios—two where there were zero bear dens in the participant's private knowledge, two where there was one bear den, and two where there were two bear dens. The other 12 scenarios where winning was impossible were removed, leaving 36 remaining.

Of the 36 remaining scenarios, the main model (with fixed parameters $\lambda_{bears} = 0.5$, $\lambda_{tracks} = 1.4$, and $cost_{lying} = 0$)[2] predicted reporting 0 bears to have the highest expected value in

[2]Since participant data had not been collected yet, the parameters of the model could not be fit. Therefore, the parameters of the model used to guide stimuli selection are different than those used in analysis.

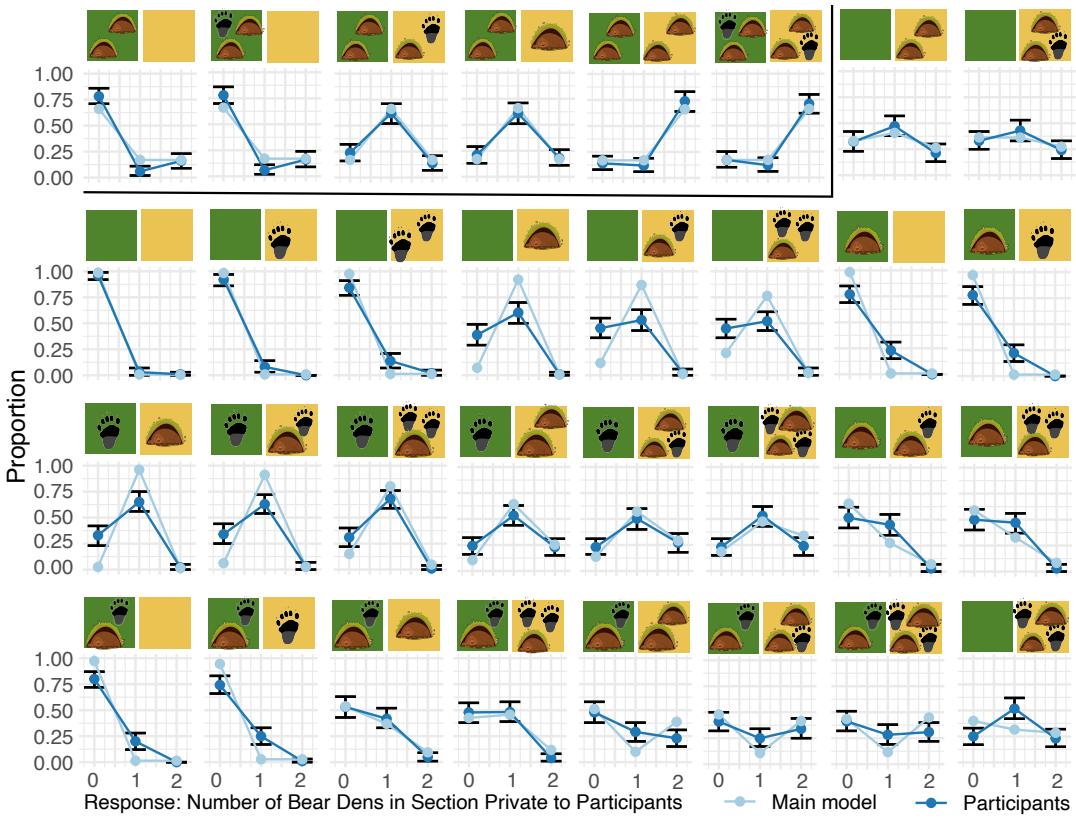

Figure 2: Trial-level results on the 32 scenarios tested. The six trials in the top row partitioned separately from the rest are the critical trials for testing lie avoidance in situations where winning is impossible. Above each graph is a small icon depicting the trial's scenario. The green area corresponds to the section of the park that the EPA surveyor and the oil company surveyed together, and the yellow area corresponds to the section that the oil company surveyed by itself. For simplicity, the area that the surveyor surveyed alone is omitted from these diagrams. Error bars represent 95% confidence intervals on the proportion of participants giving a response. The light blue points show the main model's predictions, and the dark blue points show the proportion of participants giving a response.

six of those cases. These six scenarios were selected. The remaining 30 scenarios were binned according to the difference in expected values between reporting 0 dens and reporting 1 den (the expected value of reporting 2 dens under this model was always 0), and one scenario was drawn from each bin, ensuring that these randomly-selected scenarios spanned the whole space of patterns of responses that the task might elicit. This produced a total set of 32 trials, shown in Fig. 2.

Procedure

Participants were acquainted with the rules of the *Bear and Tracks* game. The relationship between bear dens and tracks was explained as follows: bears sometimes leave tracks, and in general, the more bear dens there are in the park, the more tracks. While the presence of a track guarantees that there must be a den somewhere in the park, bears do not always leave tracks, so the presence of a den does not guarantee tracks. Participants were introduced to their role as CEO of the oil company and heavily encouraged to lie in their reports so as to build the pipeline.

All participants viewed the 32 diagrams depicting different scenarios, presented in random order. Alongside each diagram, they were asked to choose how many bear dens (among the options 0, 1, or 2) to report for the section of the park that they surveyed by themselves. After completing all trials, participants self-reported how much attention they paid during the task by using a slider from 0 to 100.

Results

Situations where winning is impossible As described in *Stimuli*, we tested six scenarios where it was impossible for participants to win (since there were two bear dens in common ground between the company and the EPA). As can be seen in Fig. 2, in these situations, the majority of participants chose to tell the truth, even though they were going to lose anyway and had nothing to gain by telling the truth. This is something that the Pathological Liar models are unable to capture, and justifies the inclusion of the $cost_{lying}$ parameter.

It could be possible that the people's cost for lying is sensitive to how far the lie is from the truth. In that case, we would

Table 1: Fitted parameters for each model. In the Pathological Liar models, $cost_{lying}$ was fixed to 0 rather than fitted.

Model	λ_{bears}	λ_{tracks}	$cost_{lying}$	τ
Main	0.4	0.5	0.34	0.25
Main x Pathological Liar	0.8	1.0	—	1.0
Shared Knowledge Only	1.1	0.1	0.30	0.14
Shared Knowledge Only x Pathological Liar	1.7	0.9	—	1.0

expect that, when the truth is two bear dens, people would prefer to report two dens, then one den, and finally, zero dens. However, those trials do not show any preference for reporting one den over zero dens. Similarly, when the truth is zero dens, we would expect to see a preference for reporting one den over two dens, but participant responses do not show this pattern. Thus, we can be confident that it is appropriate to penalize a lie categorically rather than based on its distance from the truth.

Overall results The proportion of participants giving each possible response (reporting 0, 1, or 2 bear dens) was calculated for each trial, yielding 96 data points. Then, the parameters for all four models (Main, Main x Pathological Liar, Shared Knowledge Only, and Shared Knowledge Only x Pathological Liar) were fitted via grid search so as to maximize the Pearson correlation between model predictions and participant responses. The fitted values are shown in Table 1.

The main model produced a high quantitative fit to overall participant judgments ($r = 0.90$, $CI_{95\%} : (0.85 - 0.93)$; see Fig. 3), showing that, at least at the level of aggregating over participants, this model captured participants' behavior well. Fig. 2 shows that this model was also able capture qualitative trial-by-trial patterns. However, the main model predicted that participants would increase their probabilities of lying by reporting zero bear dens as a function of the number of tracks they observed, since more tracks would indicate a higher chance of a den being present in the section that the receiver surveyed. In these cases, participants lied fairly consistently, but did not show the predicted sensitivity to the number of tracks.

Interestingly, the Main x Pathological Liar model also produced a good quantitative fit with participant responses ($r = 0.80, CI_{95\%} : (0.72 - 0.86)$). However, for the six critical trials testing for a cost of lying, it predicted that participant responses would be uniformly distributed over the three possible reports (0, 1, or 2 dens), which was not the case (as is shown in the first six trials of Fig. 2). For each report, the Main x Pathological Liar predicted 0.33, resulting in the vertical column at $x = 0.33$ in its scatterplot in Fig. 3. Overall, the Main x Pathological Liar model failed to capture participant judgments as well as the main model ($\delta = 0.10, CI_{95\%} : (0.03 - 0.17)$).[3]

Of the alternative models, the Shared Knowledge Only model performed the best ($r = 0.87$, $CI_{95\%} : (0.82 - 0.91)$) and approached the performance of the main model, but its fit was still reliably lower ($\delta = 0.03$, $CI_{95\%} : (0.01 - 0.05)$). Because this model did not consider what could be in the receiver's private knowledge, and therefore the possibility that the receiver found a bear den in that section, the Shared Knowledge Only model failed to capture participants' will-

[3] As pre-registered, confidence intervals over differences in correlations between the main and alternative models were obtained by bootstrapping over the proportion of participants responding with each possible report (i.e., the points in the scatterplots shown in Fig. 3). This ensures that the difference was not driven by an outlier.

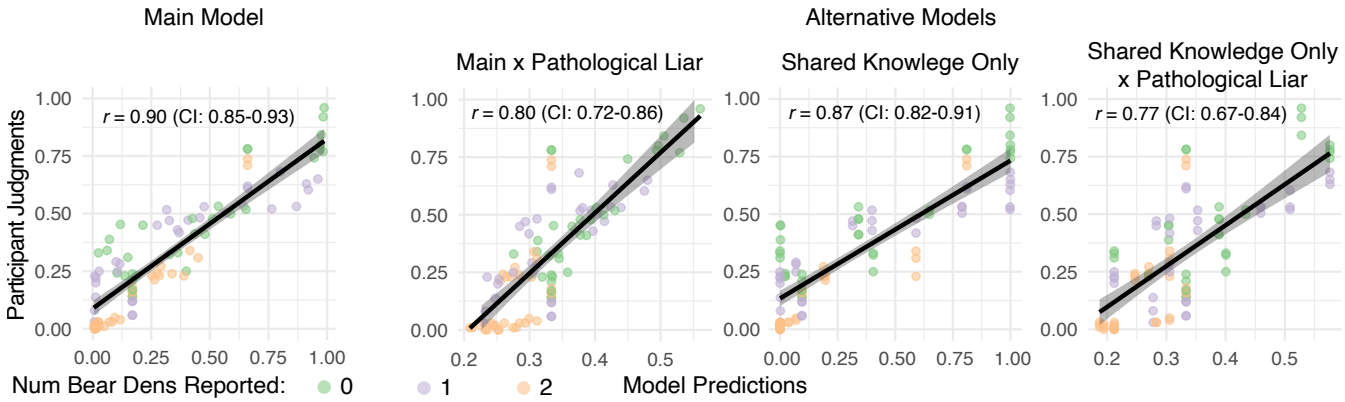

Figure 3: Scatterplots showing fits between group-level participant responses and each model. Participant responses are on the y-axis, and model predictions are on the x-axis. Each trial produces three points: the proportion of participant who reported 0 (green), the proportion of participants who reported 1 (purple), and the proportion who reported 2 bear dens (orange).

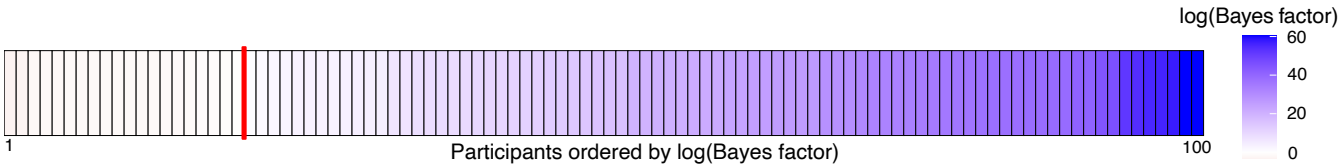

Figure 4: Depicts the natural log of the Bayes factor comparing the main and Shared Knowledge Only models. The red line partitions the 20 participants whose responses were better described by the Shared Knowledge Only model (and thus have negative log Bayes factors) from the 80 participants whose responses were better described by the main model (and thus have positive log Bayes factors).

ingness to lie and report zero bear dens on the trials where there were no bear dens in shared knowledge, and one bear den in the participant's private knowledge (see Fig. 2). Instead, on these trials, the Shared Knowledge Only model predicted that, due to the cost of lying, participants would tell the truth by reporting one bear den. This discrepancy between this model's predictions and participants' behavior produced the vertical columns of points at $x = 0$ and $x = 1$ in Fig. 3.

Finally, the Shared Knowledge Only x Pathological Liar model performed the worst, although it still captured aggregated participant responses with reasonable accuracy ($r = 0.77$, $CI_{95\%} : (0.67 - 0.84)$), especially for a simpler model that neither considered what the sender might know in private knowledge, nor a cost to lying. However, it performed substantially worse than the main model ($\delta = 0.13$, $CI_{95\%} : (0.07 - 0.21)$).

Individual-level results While our previous analyses were able to show how well our models captured aggregated participant behavior, it fails to provide insight into the models' ability to explain the responses of individual participants. To do this, we calculated the Bayes factor for each participant's responses collapsed across trials, comparing the two leading models (Main vs. Shared Knowledge Only) and assuming a uniform prior over the two models. Fig. 4 visualizes the log Bayes factor for each participant.

To assess whether individual participants' response patterns were, on average, better described by the main model, we put all the Bayes factors onto a comparable scale by inverting and negating Bayes factors smaller than 1. Averaging over these transformed Bayes factors, we found that the main model better explained individual participants' data on average ($\mu_{BF} = 2.8\text{e}24$, $CI_{95\%} : (2.9\text{e}22 - 7.0\text{e}24)$).

Not only did the main model better explain the behavior of individual participants on average, but it better explained the behavior of more participants: the responses of 80 out of the 100 participants were better explained by the main model, which is significantly different from chance based on a binomial test ($p < 1.1\text{e}{-9}$).

Discussion

We investigated different theories about how people produce strategic lies by implementing these theories as models, and testing how well they captured human lie production in a new paradigm designed to manipulate shared and private knowledge. We found that the model that considered and even inferred the recipient's private knowledge best captured human patterns of lying. This model did best when paired with a hefty penalty for lying ($cost_{lying} = 0.34$). This indicates that a lie would have to increase the chance of winning by 34% in order for participants to shift away from telling the truth.

But which exact cognitive construct this cost of lying maps onto is not quite clear. It could be that people do not expect winning through deceit to be as rewarding as winning without lying—perhaps exemplifying the hypothesis that ill-gotten gains are not as valuable (Crockett, Siegel, Kurth-Nelson, Dayan, & Dolan, 2017). Or perhaps, instead of a moral cost, this $cost_{lying}$ reflects the cognitive cost of evaluating the expected value of the lie. Perhaps people do not readily lie because of the effort required, even when the lie options are already constructed and need only to be selected.

One possible limitation of this work is that the receiver's private knowledge is conflated with the world state. It could be possible that participants consider what is in S_1 without truly representing the mind and beliefs of the receiver. This could be addressed by manipulating the task so that the receiver is always under a false belief. For example, perhaps the receiver sees double, and sees twice as many bear dens or tracks as there really are. Future experiments could incorporate scenarios where the receiver's private beliefs are false so as to test whether participants are truly inferring the receiver's private beliefs or merely the world state in the section that only the receiver can see.

Finally, this *Bears and Tracks* paradigm may also offer a fruitful domain for exploring lie detection. For example, might receivers expect senders to craft lies to be consistent with shared knowledge, but not private knowledge (especially when their private knowledge is unlikely or hard to infer)? Would that expectation influence lie detection? And are there individual differences in how people produce lies? Is that reflected in how the same individuals detect lies?

We hope that this new paradigm and formal models testing theories of lie production will prove useful in answering questions like these, advancing our scientific understanding of how humans use social cognition in adversarial settings.

Acknowledgments

We thank Maria Fernanda Rangel Carrillo, Joshua Knobe, and members of the Computational Social Cognition Lab at Yale for helpful conversations, and our CogSci reviewers for their thoughtful comments. This work was supported by NSF award BCS-2045778.

References

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), 0064.

Crockett, M. J., Siegel, J. Z., Kurth-Nelson, Z., Dayan, P., & Dolan, R. J. (2017). Moral transgressions corrupt neural representations of value. *Nature Neuroscience*, *20*(6), 879–885.

Cusumano-Towner, M. F., Saad, F. A., Lew, A. K., & Mansinghka, V. K. (2019). Gen: a general-purpose probabilistic programming system with programmable inference. In *Proceedings of the 40th acm sigplan conference on programming language design and implementation* (pp. 221–236).

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

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.

Gopnik, A., Meltzoff, A. N., & Bryant, P. (1997). *Words, thoughts, and theories.* Mit Press Cambridge, MA.

Gouzoules, H., & Gouzoules, S. (2002). Primate communication: By nature honest, or by experience wise? *International Journal of Primatology*, *23*, 821–848.

Hall, K., & Brosnan, S. F. (2017). Cooperation and deception in primates. *Infant Behavior and Development*, *48*, 38-44.

Horschler, D., Berke, M., Santos, L., & Jara-Ettinger, J. (2023). Differences between human and non-human primate theory of mind: Evidence from computational modeling. *bioRxiv*, 2023–08.

Karg, K., Schmelz, M., Call, J., & Tomasello, M. (2015). Chimpanzees strategically manipulate what others can see. *Animal Cognition*, *18*(5), 1069–1076.

Krupenye, C., & Call, J. (2019). Theory of mind in animals: Current and future directions. *Wiley Interdisciplinary Reviews: Cognitive Science*, *10*(6), e1503.

Lavoie, J. A. A., Leduc, K., Arruda, C., Crossman, A. M., & Talwar, V. (2017). Developmental profiles of children's spontaneous lie-telling behavior. *Cognitive Development*, *41*, 33-45.

Luce, R. D. (1959). On the possible psychophysical laws. *Psychological Review*, *66*(2), 81–95.

Martin, A., & Santos, L. R. (2016). What cognitive representations support primate theory of mind? *Trends in Cognitive Sciences*, *20*(5), 375–382.

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.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, *1*(4), 515–526.

Reddy, V. (2007). Getting back to the rough ground: Deception and 'social living'. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1480), 621–637.

Schmelz, M., & Call, J. (2016). The psychology of primate cooperation and competition: a call for realigning research agendas. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(20150067).

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.

Verschuere, B., Köbis, N. C., Bereby-Meyer, Y., Rand, D., & Shalvi, S. (2018). Taxing the brain to uncover lying? meta-analyzing the effect of imposing cognitive load on the reaction-time costs of lying. *Journal of Applied Research in Memory and Cognition*, *7*(3), 462–469.

Wellman, H. M. (2014). *Making minds: How theory of mind develops*. Oxford University Press.