Designing strategic false utterances: Recursive social inference in lying and lie detection

Lauren A. Oey (loey@ucsd.edu)
University of California, San Diego, Department of Psychology

The deluge of fake news stories in recent years has sparked an interest in understanding why people are susceptible to misinformation, and why some lies are more effective deceptions than others (Rubin, Chen, & Conroy, 2015). Here, we aim to formalize the cognitive processes underlying the strategic behavior of liars and lie detectors in deceptive communication. We argue that skillful and undetected lying requires a rational theory-of-mind that considers the feasibility of the lie from the interlocutor's perspective (e.g. "I know this is a lie, but would you suspect?").

To examine how humans behave in deceptive situations, we introduced a novel dyadic lying game that allows for quantitative measurement of lies. In each round, one player draws a sample of 10 marbles from a box of red and blue marbles, of which k are red. Both players see the distribution of red-to-blue marbles (the contents of the box); however, only the marble-drawer sees the sample of red marbles k. The marble-drawer then reports a number k^*, and can choose to (a) tell the truth, or (b) lie about the number of red marbles drawn. The other player then responds to the marble-drawer's reported value by choosing to either (A) accept the reported value as truth, or (B) reject the value as being a lie (call 'BS'). Both players' payoffs are determined by the marble-drawer's and the responder's decisions: it is advantageous for the marble-drawer to lie but not get caught; meanwhile, the responder seeks to catch the marble-drawer in a lie but avoid false accusations.

To test the hypothesis that skillful lying requires a recursive theory-of-mind, we developed a recursive adversarial Bayesian model that aims to formalize the latent inferential processes involved. By this model, the liar's decision to choose a particular utterance (either lie or truth) depends both on the true state of the world and their inference about whether the lie detector would be likely to call out that utterance as a lie. Likewise, the lie-detector chooses whether to make a lie accusation both on the prior expectations about the statistics of the world, and on their model of how the liar might act in different world states. We compared the Recursive Theory-of-Mind model to (1) a No-Theory-of-Mind model--a model that does not perform inference on the other agent and (2) an Oracle model--a model that has perfect knowledge about the other agent's behavior.

Across three sets of experiments, we compared human performance to the model predictions. All participants played against an AI across multiple rounds of the game, with both players alternating between the marble-drawer and responder roles. In Experiment 1 (n=193), we found that humans lied in a manner qualitatively predicted by the Recursive Theory-of-Mind model: they reported values in a non-linear drift pattern relative to k, with more lies when k was below, and more truths when k was above, a drift point around the expected mean (5). In addition, responders called BS in a sigmoidal trend relative to the liar's reported k^*, where higher k^* values were rejected more often and lower k^* values were rejected less. These results are consistent with the predictions of the Recursive Theory-of-Mind model. In Experiment 2 and 3, we asked whether participants could rationally integrate information about changes to the true distribution of the world and to the other player's goals, respectively, into their judgments about realistic lies. We found that across both manipulations participants adjusted their lies and detection in the direction predicted by the Recursive Theory-of-Mind model.

Overall, we developed a recursive Bayesian model of the cognitive processes of agents in deceptive situations and found that our model qualitatively predicted the pattern of human performance across three experiments. Our formalization and behavioral experiments have allowed us to start to characterize how liars and lie detectors behave rationally in determining what lies to tell and which lies to call out.