Increased Vigilance in Monitoring Others' Mental States During Deception

Nicholas D. Duran (nduran2@ucmerced.edu) **Rick Dale (rdale@ucmerced.edu)**
Cognition and Information Sciences, The University of California, Merced, Merced, CA, 95343

Abstract

Communicating false information during social interactions places unique cognitive demands on the speaker. One such demand is the increased need to track what another knows to avoid introducing contradictory information. The present study examines the minimal conditions required for vigilant mental state monitoring by using a game-like task that required participants to respond to virtual partners' questions with true or false information. In this task, there were no explicit demands to engage in mental state monitoring. We found increased response times when the answers potentially violated a partner's belief states - but only if participants believed their partner to be an actual cognitive agent. These effects were also shown to be additive to the simple demands in suppressing a truth bias while responding falsely. We argue that participants exert greater cognitive effort as an automatic response elicited by being situated in an interactive, deceptive context.

Keywords: deception; belief monitoring; self; other

Introduction

In this paper, we examine the underlying cognitive processes that are involved in deceptive responding during a real-time social interaction. We propose two primary processes: 1) the inhibition of a prepotent truth response while responding falsely, and 2) the need to track the knowledge of the person to whom one is falsely responding. Both of these processes are commonly associated with executive control and working memory, and thus are hypothesized to involve considerable cognitive costs. Previous studies on deceptive behavior have examined these processes in isolation. In this paper, we show the integrated effects of each process in a single task. Moreover, we provide evidence that when participants merely know that they may have to convey false information, even when the risk of detection is negligible, it is sufficient to induce vigilant tracking of the knowledge of the false information's recipient. However, as an important caveat, this effect only occurs when a person takes an *intentional stance* toward another (Dennett, 1987). As we argue below, such tracking is likely an adaptive response that ensures false or true information conveyed throughout an interaction remains consistent, thereby minimizing the risk of detection.

Background

Human interaction often involves the exchange of deceptive information. Although these fabrications are mostly innocuous, as in lying to boost one's own credentials or to protect another's feelings, they are routinely introduced into conversations (DePaulo & Kashy, 1998). Nevertheless, the risk of detection when communicating false information is always present. To minimize this risk, plausibility must be maintained, and and as such, sophisticated cognitive control processes are likely recruited. A primary need is the suppression of a truth bias while responding falsely (Duran, Dale, & McNamara, 2010), as well as a need to remember what information is false and maintain this falsity as the ostensible truth in the mind of another (Sip, Roepstorff, McGregor, & Frith, 2008). A number of neuroimaging and reaction time studies have explored truth suppression, showing increased executive function processes involving the inhibition of predominant responses and the resistance to interference (see Spence et al., 2004, for a review).

Much less is known, however, about the processes underlying real-time deception in social interactions. Some work has focused on the recipients of deception and their beliefs about the hidden motivations of their partners (e.g. Schul, Mayo, & Burnstein, 2004). Conversely, others have looked at those doing the deception and have found evidence for active monitoring of their partners' suspicion, purportedly for the purpose of manipulating others' beliefs about their own goals and intentions (Bhatt, Lohrenz, Camerer, & Montague, 2010; Carrion, Keenan, & Sebanz, 2010). In these studies, the modeling of another's mental state consists of the impressions that the other is likely to form. Participants are given opportunities to mislead a partner about the true nature of some privileged knowledge, as a poker player does when bluffing about the cards in their hand. Successful participants are those that strategically fend off another's suspicion with well-timed true responses. These true responses, construed as "deceptive truth," are believed to result from mental state monitoring, and invoke neural regions of cognitive effort similar to that associated with various theory of mind tasks. Beyond considerations about the general impressions or motivations another might possess, deception likely involves the specific content one believes another to believe about events and objects in the world following the deceptive act.

We explore this aspect of mental state monitoring in deceptive interaction; specifically, the need to encode and maintain the factual content of what another might know. This is particularly difficult during an extended interaction where new information is introduced and must be integrated into deceiver's knowledge of another's supposed knowledge. If one is to give another false information about their whereabouts, such as saying they were not at the local Sears on Saturday, the deceiver needs to maintain this model of events for the other, and apply it consistently downstream in the conversation to avoid saying "yes" to a question like, "Were you shopping on Saturday?" From this example, implications that might be drawn from the earlier presented false information also need to be maintained by the deceiver, such as knowing that being at Sears implies shopping. Such a model of others' belief states is likely an adaptive evolutionary response that minimizes the risk of detection (Premack, 2007), but one that also enacts considerable cognitive costs that combines with those associated with truth suppression (a process that can occur regardless of an audience). To evaluate both these processes in an integrated manner, we turn to a novel *guessing game* task.

The basic structure of this task involves a partner attempting to guess the identity of an object of which another partner is solely aware. This task is similar to the game of "Twenty Questions," where a "questioner" tries to guess a person, place, or thing by asking yes or no questions. After the allotted number of questions have been asked, the questioner must wager a guess. This task presents a situation where the mental state of another (i.e., the questioner) incrementally converges on what another has in mind (i.e., "answerer"). Although the uninformed questioner is making explicit attempts to converge, the same does not necessarily hold for the partner who is answering yes or no. That is, to succeed in this task, the answerer does not need to track the evolving image emerging in the questioner's mind. However, in our version of the Twenty Questions game, the answerer is told that there is the possibility of having to give the questioner false information. This simple instructional manipulation is hypothesized to elicit increased vigilance in tracking the mental state of the questioner, largely because the answerer now has to maintain the veracity of what the other believes. Important to the task set-up, there are no explicit instructions for monitoring another's mental state, nor does it affect the success of completing the task.

Mechanisms involved in such a situation are likely automatic and triggered in response to particular contexts (German, Niehaus, Roarty, Giesbrecht, & Miller, 2004) - one of which, as we argue, is the communication of false information. However, merely being situated in a context that requires transmission of false information is not sufficient. The participant must also take an intentional stance toward their partner, a partner who is thought to have their own set of beliefs, desires, and knowledge states. In other words, participants must consider their partners as having minds worth tracking (Gallagher, Jack, Roepstorff, & Frith, 2002). One test of this claim is to compare participants (i.e., answerers) who believe their partner to be real versus a computer simulation, while holding all other features of the interaction equivalent. Accordingly, on critical trials all participants should find the false responses more challenging than the truth (due to suppression of a truth bias); however, for those who take an intentional stance, they will experience added difficulty because they will also be violating knowledge their partner possesses.

In what follows, we describe in greater detail the method used to assess other-directed mental state monitoring during deception. We then report the findings from two experiments that incorporate crowdsourcing techniques. We end with a brief discussion of the findings and limitations.

Mental State Monitoring in a Guessing Game

Initial Set-up

The current task was implemented as an online, Flash-based game that makes use of key features in Amazon Mechanical Turk (AMT). AMT is a crowdsourcing platform that allows participants (i.e., "Workers") to sign up to complete tasks posted by other users (i.e., "Requesters"). There are many advantages of using crowdsourcing techniques (see Munro et al., 2010, for a review), with one notable advantage being the ability to create an illusion of connectivity, whereby simulated, recorded partners can act as convincing interactive partners (Duran, Dale, & Kreuz, 2011). To achieve this illusion in this study, participants were recruited under the pretext of examining "how people solve problems while receiving misleading information." They were then told that the task involved beta software that allows us, the Requesters, to connect two Mechanical Turk Workers, but with software that only allows a one-way transmission of audio. The participant's partner (a recorded simulation of a male's voice) was always designated as the role of questioner and was the one who would transmit audio. A series of validity checks were then presented that highlighted the connection, such as a "connection screen" where participants ostensibly waited for the software to locate and connect them to their partner, and an "introduction screen" where the recorded partner introduced himself and provided a secret codeword for the participant to enter, ostensibly verifying to the recorded partner that they were "connected" to a real person. As described further below, checks were used to ensure participants believed this sham connection.

Basic Game Structure

In the initial instructions, participants were provided a demonstration of how the task was structured (see Figure 1 for a flow diagram). First, participants were told that they would be presented with one of two objects (an alarm clock or a red apple) that only they could see. Their partner would then ask a yes or no question to attempt to guess the identity of the object, and once the question was asked, they would trigger a "GO" button that would appear at the bottom-center position of the participant's screen. When the participant clicked this button, a response screen would appear with "YES" and "NO" response buttons positioned in the top-right and top-left corners. On this screen, a prompt to respond with a "TRUTH" or "LIE" also appeared in the middle of the screen.[1] Thus, if the partner had asked, "Is it a person?," the participant had to navigate their computer mouse from the bottom of the screen to the "YES" response button. The response was then transmitted to the partner and a short pause was introduced to allow the partner to "formulate" another question before the next trial began.

[1] The use of a "LIE" response prompt has been used in previous research to approximate deceptive behavior and has been shown to invoke similar physiological and neurological reactions as unsolicited deception. However, we acknowledge that this is still an approximation of deceptive behavior, which we address in Study 2.

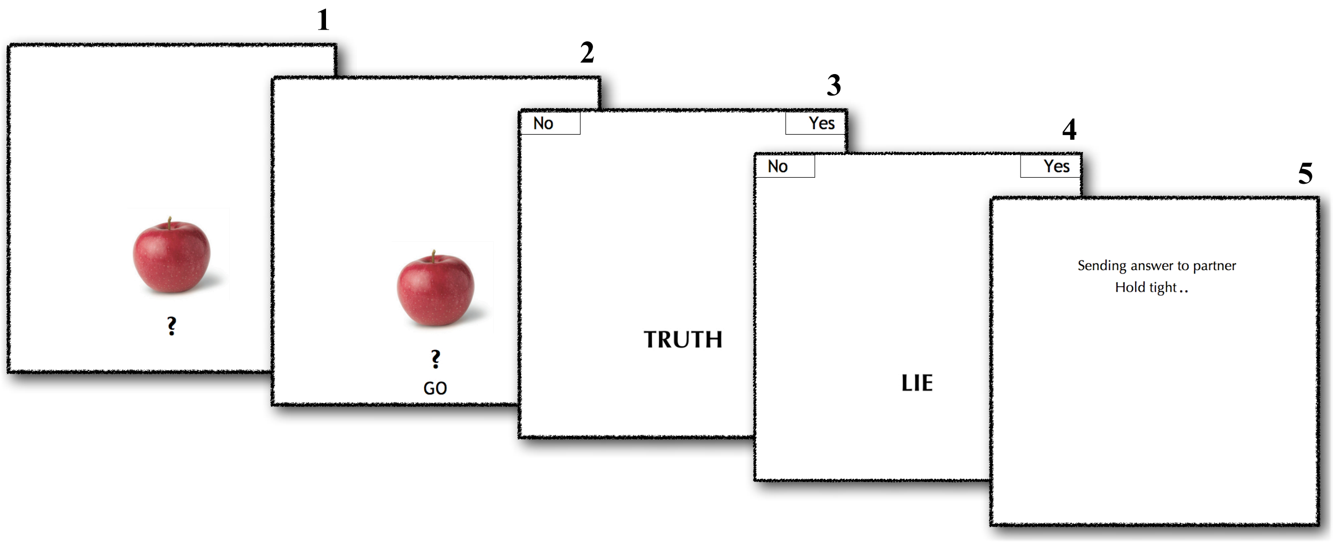

Figure 1: Flow of interaction in experimental task. In (1), after initial instructions, a typical trial begins by presentation of the to-be-guessed object. The partner then asks a question about the identity of the object; (2) After the question is asked, "GO" appears and participants click to respond; (3) and (4) Participants are given a prompt to respond truthfully or falsely, and do so by clicking "YES" or "NO" boxes; (5) Response is "sent" to partner.

Monitoring and Contradicting Given Information

Two rounds of the guessing game were played, counterbalanced across the object to be guessed (clock or apple). In each round, the partner asked 5 questions before they made a guess.[2] There were 16 variations of each question set for each object, with each set randomly selected for each participant. In both rounds, participants were prompted to answer truthfully for the first three questions, but for the 4th or 5th questions, they were prompted to respond falsely. These falsification questions presented a situation where participants might violate their partner's possible belief states. What this means is that by the 4th or 5th questions, the partner has eliminated numerous options of what the object might be, allowing a greater possibility for some objects, like an alarm clock, to be mentally represented. If participants take into consideration this reduced mental set of their partner, a response that denies the possibility of an alarm clock acts to contradict what the partner was increasingly likely to believe. This type of *general* violation constituted one round of the guessing game.

In another round, a *specific* violation was presented. In the corresponding scenario, the question being falsified contradicted information explicitly given earlier in the informational exchange. For example, if an early question like, "Can it be used on a daily basis?," is initially confirmed by the participant as being true, the participant can infer that their partner believes that the unknown object is "common." When a later-occurring question such as, "Is it fairly common?," must then be falsified, the participant is now violating the more specific belief their partner was inferred to have. An example sequence of questions is given in Table 1. These two types of violation, general and specific, were presented to the participant across the two rounds, in counterbalanced order.

Question Sequence	*Expected Response*	*Prompt*
1. Is it a person?	No	True
2. Did someone invent it?	**No**	**True**
3. Do people eat it?	Yes	True
4. Does it grow underground?	No	True
5. **It is a man-made product?**	**Yes**	**False**

Table 1: Sample questions asked by simulated partner who is attempting to "guess" an apple. The later-occurring Question 4, in which participants are prompted to falsify, contradicts specific information given earlier in Question 2. Questions were also counterbalanced for whether critical trials required a "yes" or "no" response.

Establishing Intentional Stance

To examine the incurred processing costs of violating the mental state of another, we compared participants who believed that they were interacting with a real partner versus those who did not. Participants who did not believe they were interacting with a real person were hypothesized to be less likely to attribute mental states to the simulated partner, and

[2] Limited to 5 questions to avoid memory interference or decay that would have likely resulted from 20 questions.

thus should not experience the associated processing costs on the critical false trials. The only costs these participants should experience is that attributable to the suppression of a truth bias when responding falsely. To identify participants who did or did not believe they were interacting with a real partner, we asked two critical follow-up questions at the end of the task. The first probed whether the participant would give a small monetary bonus (paid by us, the Requesters) to their partner for the quality of questions asked. The rationale for including this question is that a participant who suspects that their partner is a mere recording is unlikely to give a reward. The second question was more direct, and asked participants to rate on a scale from 1 to 7 the degree to which they thought they were connected to a real person. Participants who would give a reward, and were on the upper end of the scale for their belief that they were interacting with a real person, were considered those who would take an intentional stance.

Experiment 1

We collected data from 104 participants. One subject was excluded for answering two of the 10 questions incorrectly. We also removed excessively long trials that were over 8000 ms (0.73% of data), and from this truncated set, we removed trials that were more than three SDs (855 ms) above the mean response time (2308 ms). This resulted in a loss of 2.48% of the data. Forty-eight participants also self-selected into a group who believed that their partner was real, with the remaining participants believing that their partner was not real.

Results

A mixed effects ANOVA was used to compare the difference in response times for the two groups of self-selected participants, with a within-subjects factor of whether a trial required a true or false prompt[3] [4] Subject was entered as a random effect. The analysis was conducted using the lmer package in the R statistical software. In this package, *p*-values are computed with 10,000 Monte Carlo Markov Chain simulations, using lmer's pvals.fnc function (Baayen, Davidson, & Bates, 2008). We report these *p* values and the unstandardized effect estimates for the main effects and interactions.

The results indicate a main effect for participant type (those who believe vs. not believe), B = 342 ms, p = .003; as well as for prompt type (false vs. true), B = 481ms, $p < .001$. There was also an interaction between participant and prompt type, B = 277 ms, p = .05. In follow-up tests to examine this interaction, it appears that for both believers and non-believers, the false response critical trials showed greater response latencies than true response latencies: believers, B = 622 ms, $p < .001$; and non-believers, B = 345 ms, p =. 003. This effect corresponds to the greater cognitive difficulty associated with suppressing a false response. Moreover, the magnitude of the false response time latencies for believers was much greater that those who did not believe, B = 500 ms, $p < .001$ (see Figure 2). This finding suggests that believers noticed a contradiction forced by the false prompt that the non-believers did not. The likely reason is that believers had an active model of the other's mental state and experienced greater cognitive effort in consulting and ultimately violating this knowledge.

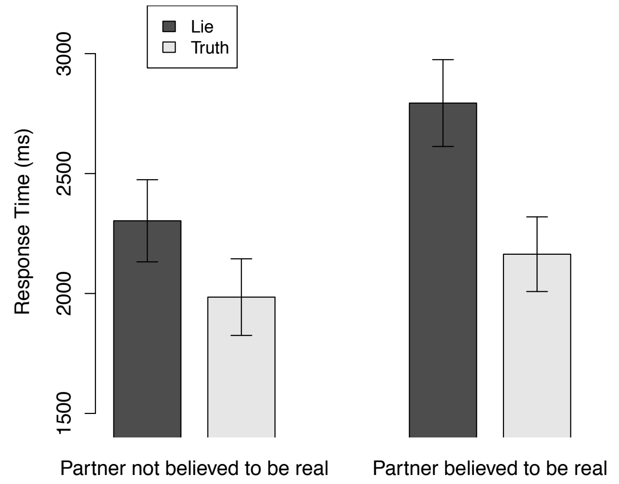

Figure 2: Overall, the response time latencies are much higher for false (lie) responses compared to true responses, with the greatest response latencies for participants who believed their partner to be real. There are no significant differences across true responses.

Experiment 2

In Experiment 1, participants appear to use true knowledge provided earlier *upstream* in an interaction to influence later *downstream* responses. As a further test, we examined whether participants encode and consistently maintain false information that is provided upstream. The set-up of this task is such that participants should respond falsely to a later-trial critical question to verify the accuracy of earlier presented false information. To do so, participants are allowed to freely respond without prompts. We hypothesize that participants who believe they are interacting with a real partner will be more likely to choose to respond falsely. In other words, they are more likely to respond falsely *again* to preserve the consistency of their partner's beliefs (because of false information they presented upstream). This response will enact similar cognitive demands as evidenced in Experiment 1.

Modified Method

Participants first provided their partner with a particular belief about the to-be-guessed object early in the interaction.

[3] Because of the smaller number of false critical trials compared to true trials, and because the false trials always occurred near the end of the interaction, we only analyzed the true trials that occurred immediately before the presentation of the false prompt trials. The inclusion of all true trials does not radically alter the reported findings, as the response pattern in the data remains consistent. However, by doing so, the significant interaction is now only marginally significant ($p = .10$).

[4] The comparison between the false prompts that draw on general and specific belief violations showed no statistically significant differences, and thus are combined into one condition.

This is done by prompting the participant to respond falsely on the second question asked by their partner. Later, downstream in the questioning (on the critical fourth question in the questioning sequence), participants are allowed to choose whether to respond falsely or tell the truth. In one round of the guessing game, the downstream question is related to the earlier false information, and thus if a participant is monitoring what the other knows, and wants to stay consistent with the false depiction of the object, they will voluntarily respond falsely (see Table 2 for a sample sequence of questions). By consulting and acting on the knowledge state of the other, we expect this to incur a processing cost. As a type of control, in a second round of the guessing game, the downstream question is unrelated to the earlier false information, and thus there is no explicit need to consult what the other knows on the critical question. Thus, the differences in the *information related* round should no longer be present in this *information unrelated* round.

Question Sequence	*Expected Response*	*Prompt*
1. Is it a thing	Yes	True
2. Can you easily pick it up?	**No**	**False**
3. Is it found in people's homes?	Yes	True
4. Can it be moved?	**?**	**--**
5. Is it something people might use on a daily basis	?	--

Table 2: Sample question set asked by simulated partner who is attempting to "guess" an alarm clock. The downstream Question 4, which requires a free response, relates to information falsified in Question 2. There are 8 variations of question sets for each object (total of 16).

Results

One hundred and seventy-six participants supplied data via Mechanical Turk. Five participants were removed for answering two or more of the 10 question incorrectly, and an additional 16 were removed for failing to provide at least one false response in the final unprompted questions. Outlier trials, those trials that exceeded 3 SDs above the mean were also removed. Furthermore, based on follow-up questions, 80 participants were self-selected as those who believed they were interacting with a real partner, 71 believed they were interacting with a simulation, and four participants' beliefs were undetermined. Thus, 151 participants provided data where their false and true responses could be evaluated.

The main analysis specifically targets the round in the guessing game where there was an opportunity to maintain the false information introduced earlier in the question sequence (i.e., "information related" round; see Figure 3a). By freely answering false on the fourth (critical) question in the sequence, the false belief state of the questioner is maintained. We hypothesized that this process requires greater processing time because the deceiver must consult what the other knows, recognizing that to respond true would elicit a contradiction. Importantly, such behavior is likely to occur only when a participant believes they are interacting with a real partner. We found evidence for this hypothesis in a simple t-test evaluating the critical false response trials between participants who did or did not believe they were interacting with a real partner, t(76) = 2.20, *p* = .03 (Figure 3a).

It should also be noted that the false response trials in the above analysis represented responses from 41 of the 80 participants who could be classified as believers. This number is fewer than expected given the hypothesis of increased vigilance in maintaining the partner's belief states. However, when participants who believed they were interacting with a real partner did answer truthfully (thereby contradicting a partner's mental state), these response times were the second highest of all trial groups (Figure 3a). These elevated scores suggest that participants are aware, at some level, that they are violating their partner's mental state. This is supported by a significant main effect in a mixed effects ANOVA for participant type (belief vs. not belief), showing that participants who believed they were interacting with another, despite answering truthfully or falsely, had increased response times compared to those who did not believe, *B* = 621 ms, *p* = .02.

Finally, as hypothesized for the round in the guessing game where the upstream information was unrelated to the downstream information, no differences were found between participants who did or did not believe they were interacting with a real partner (see Figure 3b)

General Discussion

Across two experiments we examined response behavior in a context where participants had to transmit false information to another. We found evidence that participants experience greater cognitive effort in suppressing a truth bias; and furthermore, show evidence of increased effort when they are confronted with a response that violates or potentially violates the mental state of another (as measured by response latencies). For the latter, we argued this increased effort results from an active, or vigilant, monitoring of what another believes. Participants appear to do so as long as they think they are interacting with a "mindful" agent, and also do so despite instructions that have no explicit requirement to consider others' mental states.

A limitation in this study, found in Experiment 2, is that participants who believed their partner to be real did not overwhelmingly choose to maintain the false information provided upstream in the interaction. One reason is that being detected as providing contradictory information carried little consequence, thus there was little motivation to choose a false response. Other research also suggests strong individual differences in whether participants consider their partners' sus-

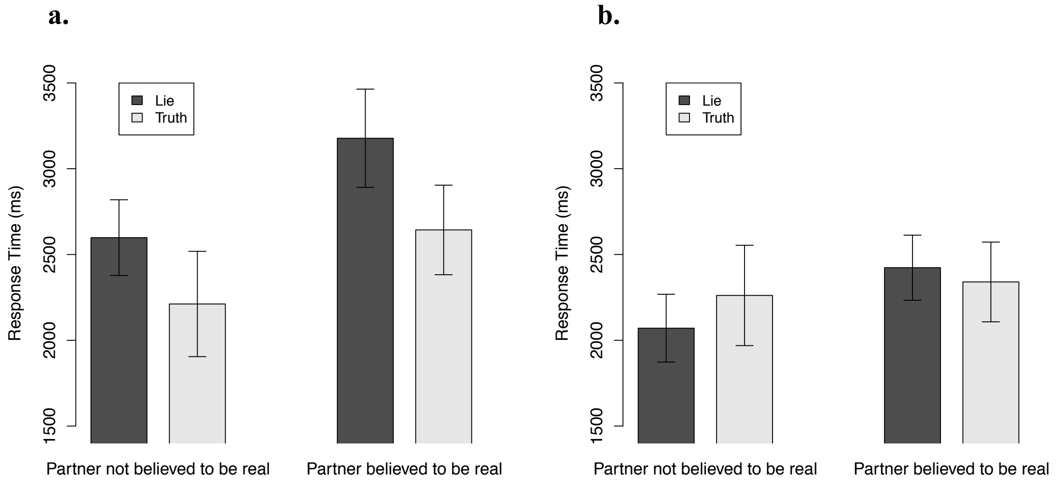

Figure 3: In (a), response times for participants who answered falsely on the critical free response questions is highest for those who believed they were interacting with a real partner versus those who did not believe. There are no differences for true responses. In (b), the control with unrelated information downstream showed no significant effects.

picion in low risk interactions (Bhatt et al., 2010). Despite this limitation, those participants who did think they were interacting with another, and who responded falsely, were the only group to be influenced by the contradictory information in the information stream.

In sum, we explored how truth biases and social factors are jointly involved in simulated "deceptive acts." While we grant that these are basic cognitive experiments that can only loosely approximate naturalistic contexts, we would argue that the approach opens new avenues of investigation. The underlying *cognitive mechanisms* of deception are still being sought. By employing basic cognitive experimentation in simple but controllable tasks, we could gain a more systematic understanding of the mechanisms underlying deceptive acts. These experiments are a step in that direction.

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390-412.

Bhatt, M. A., Lohrenz, T., Camerer, C. F., & Montague, P. R. (2010). Neural signatures of strategic types in a two-person bargaining game. *Proceedings of the National Academy of Sciences*, *107*, 19720.

Carrion, R. E., Keenan, J. P., & Sebanz, N. (2010). A truth that's told with bad intent: An ERP study of deception. *Cognition*, *114*, 105-110.

Dennett, D. C. (1987). *The Intentional Stance*. Cambridge, MA: The MIT Press.

DePaulo, B. M., & Kashy, D. A. (1998). Everyday lies in close and casual relationships. *Journal of Personality and Social Psychology*, *74*, 63-79.

Duran, N. D., Dale, R., & Kreuz, R. J. (2011). Listeners invest in an assumed other's perspective despite cognitive cost. *Cognition*.

Duran, N. D., Dale, R., & McNamara, D. S. (2010). The action dynamics of overcoming the truth. *Psychonomic Bulletin & Review*, *17*, 486-491.

Gallagher, H. L., Jack, A. I., Roepstorff, A., & Frith, C. D. (2002). Imaging the intentional stance in a competitive game. *Neuroimage*, *16*, 814-821.

German, T. P., Niehaus, J. L., Roarty, M. P., Giesbrecht, B., & Miller, M. B. (2004). Neural correlates of detecting pretense: Automatic engagement of the intentional stance under covert conditions. *Journal of Cognitive Neuroscience*, *16*, 1805-1817.

Munro, R., Bethard, S., Kuperman, V., Lai, V., Melnick, R., Potts, C., et al. (2010). Crowdsourcing and language studies: the new generation of linguistic data. *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 122-130.

Premack, D. (2007). Human and animal cognition: Continuity and discontinuity. *Proceedings of the National Academy of Sciences*, *104*, 13861.

Schul, Y., Mayo, R., & Burnstein, E. (2004). Encoding under trust and distrust: The spontaneous activation of incongruent cognitions. *Journal of Personality and Social Psychology*, *86*, 668-679.

Sip, K. E., Roepstorff, A., McGregor, W., & Frith, C. D. (2008). Detecting deception: The scope and limits. *Trends in Cognitive Sciences*, *12*, 48-53.

Spence, S. A., Hunter, M. D., Farrow, T. F. D., Green, R. D., Leung, D. H., Hughes, C. J., et al. (2004). A cognitive neurobiological account of deception: Evidence from functional neuroimaging. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *359*, 1755-1762.