Predictive Models of Malicious Behavior in Human Negotiations

Zahra Nazari and Jonathan Gratch
Institute for Creative Tachnologies
Univeristy of Southern California, Los Angeles California
{zahra, gratch}@ict.usc.edu

Abstract

Human and artificial negotiators must exchange information to find efficient negotiated agreements, but malicious actors could use deception to gain unfair advantage. The *misrepresentation game* is a game-theoretic formulation of how deceptive actors could gain disproportionate rewards while seeming honest and fair. Previous research proposed a solution to this game but this required restrictive assumptions that might render it inapplicable to real-world settings. Here we evaluate the formalism against a large corpus of human face-to-face negotiations. We confirm that the model captures how dishonest human negotiators win while seeming fair, even in unstructured negotiations. We also show that deceptive negotiators give-off signals of their malicious behavior, providing the opportunity for algorithms to detect and defeat this malicious tactic.

1 Introduction

Negotiation is an important focus of artificial intelligence research [Jennings et al. 2001, Fatima et al. 2004, Kraus et al. 2008, Traum et al. 2008]. While research has emphasized fully-automated negotiations, recent efforts have sought to analyze or engage with human negotiators [Hindriks and Jonker 2008, Rosenfeld et al. 2014]. Unlike rational frameworks where “talk is cheap,” information exchange is crucial in human negotiations. By sharing information, negotiators form better models of their opponent, gain insight into the joint structure of the task, and reach more efficient solutions that benefit all. However, negotiators could subvert this process for selfish gain. Recently, we introduced *the misrepresentation game*, a game-theoretic analysis of how deceptive agents could manipulate information exchange to gain disproportionate rewards while seeming honest and fair [Gratch et al. 2016]. We presented a provably-optimal solution to the misrepresentation game, and provided empirical evidence that the solution works well against human opponents. Though promising, these results come with several qualifications. Our proof of optimality and our experimental evaluation rely on strong assumptions that are unlikely to hold in practice, thereby undermining the generality of the findings. Here, we seek to correct these limitations. We evaluate the formalism without these restrictive assumptions using a large corpus of human face-to-face negotiations. We confirm that the model is consistent with how dishonest human negotiators gain advantage in these unstructured negotiations. Specifically, deceptive negotiators lie in the ways predicted by the model, and these liars significantly out-performed their honest counterparts. Fortunately, we also show that people give-off signals of their own deception, providing the opportunity for algorithms to detect and defeat this malicious tactic.

2 Background

In bilateral negotiations, parties seek to reach a joint agreement over several issues. For example, in salary negotiations, parties are not only interested in salary but also benefits. Negotiators cannot be purely self-interested. Each must consider the interests of the other in order to reach an agreement at all. Further, by carefully considering each other’s interests, parties can often find solutions that benefit both sides. This is complicated, however, in that negotiators often do not know the preferences of their partner, but can only infer them through exchange of information or exchange of offers.

Discovering what one’s opponent wants is a major challenge for human and automated negotiators alike [Hindriks and Tykhonov 2008]. Parties may be reluctant to share information, but with accurate models, negotiators can find mutually beneficial solutions. For example, an employer might only want to pay a part-time salary but assumes a prospective employee wants to work full time to obtain health benefits. If the employee reveals they already receive health benefits through their spouse and would prefer a flexible schedule instead, both sides can get what they want. Negotiations where parties have complementary preferences are known as *integrative* negotiations (in contrast, *distributive* or “fixed-pie” negotiations arise when parties have competing interests). Indeed, research shows that human negotiators generally obtain better outcomes in integrative settings when they honestly exchange preference information [Thompson 1991].

Unfortunately, information exchange can be exploited by malicious actors. If one side has an information advantage (i.e., they obtain a good model of their opponent without revealing their own preferences), they can use deception to gain an unfair outcome while seeming to be fair. One approach is to feign high interest in a low-value issue; sometimes called a “bogey tactic” [Schweitzer et al. 2002]. For example, the employer could deceptively argue that they need a full-time employee but would be willing to accept a flexible schedule for a much-reduced salary. Human negotiators often freely and honestly exchange information [Thompson 1991], but some clearly exploit this cooperative tendency for competitive advantage [O'Connor and Carnevale 1997].

Assuming a negotiator (human or software) is willing to engage in deception, they must reason about how best to lie. Recently, we provided a game-theoretic analysis of this problem and proposed a solution [Gratch et al. 2016]. In what we call the *misrepresentation game*, a malicious actor is faced with the challenge of how to gain an information advantage and then to misrepresent their own preferences to win. The game builds upon and formalizes deceptive tactics found in the negotiation literature (e.g., [O'Connor and Carnevale 1997]). Solutions to this game could be used to guide deceptive agents, but also to help agents or human negotiations understand when they are being deceived.

3 The Misrepresentation Game

The misrepresentation game adopts a standard formulation of negotiation known as the multi-issue bargaining task. This requires parties to reach a joint agreement over a set of issues, where each issue can have one of many possible levels. For instance, in a job negotiation, negotiators must compromise on the level of the salary (a continuous variable) and the number of vacation days (an integer variable). Each party’s proclivity for a given deal can be represented by a utility function that returns a utility given some level for each issue. For example, Figure 1 illustrates a negotiation where the utility for each party is a linear combination of the number of items of fruit received, weighted by the unit-price of each fruit.

The misrepresentation game differs from standard game-theoretic analysis – where self-interested rational actors interact – and also from behavioral game theory – which seeks to explain how actual humans, bound by cognitive biases and social conventions like honesty and fairness, behave in practice [e.g., Colman 2003]. Rather, it considers how a rational self-interested actor can benefit by an awareness of human biases and conventions. Such an actor might need to appear honest and fair, but need not be bound by these rules. Real-world examples of this include people with psychopathy or a Machiavellian personality [Paulhus and Williams 2002]. In the future, it may also include malevolent software agents.

Following Nash’s common-sense axioms of cooperative behavior [Nash Jr 1950], pro-social negotiators prefer deals that are efficient and fair. A deal is efficient if it is impossible to make one party better off without making the other worse, and such deals can be discarded from consideration as they can always be improved for one party without harming the other. A deal is fair if it minimizes the difference in utility received by the two parties. Any deal that satisfies these axioms (there can be more than one) is known as the *Nash bargaining solution*. In the case of perfect information, human negotiators tend to prefer the Nash solution. Fairness concerns, in particular, loom large even when significant stakes are involved (e.g., see [Fehr and Schmidt 1999]).

With these preliminaries, the misrepresentation game requires the deceiver to claim a false set of preferences such that a personally-advantageous solution appears, to the opponent, to be a fair and efficient solution. In [Gratch et al. 2016], we formalized the problem as follows. Let a_i be the deceiver’s true preferences for each issue *i*, b_i the true preference of the opponent, and x_i and y_i represent the agreed upon levels for each player (e.g., salary or number of apples). The objective is to find the optimal ‘false’ preference $\bar{a}_i$ such that maximizes the deceiver’s utility, given that the negotiated solution appears to be efficient and fair if $\bar{a}_i$ is believed. This can be formalized as an optimization problem (under the assumption that issues are independent) as:

$$\max_{\bar{a}_i} \sum_{i=1}^{n} a_i \,.\, x_i \qquad (1)$$

where x_i is the optimal solution of

$$\max_{x_t, y_t} \bar{a}_i \,.\, x_i \qquad (2)$$

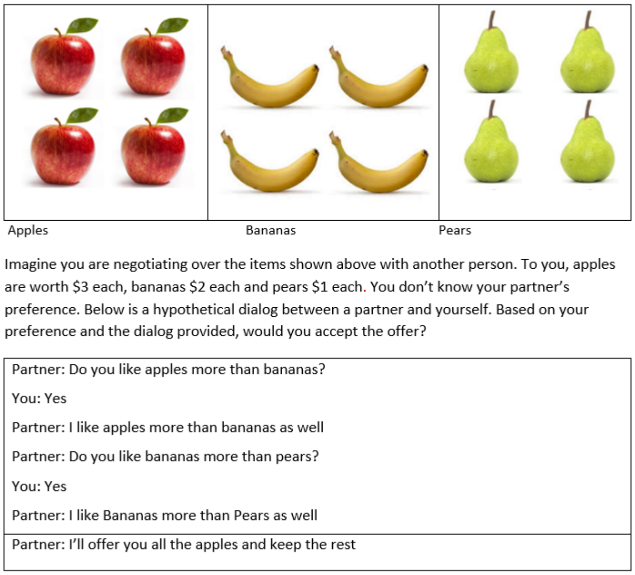
such that :

$$\sum_{i=1}^{n} \bar{a}_i \,.\, x_i \leq \sum_{i=1}^{n} b_i \,.\, y_i \qquad (3)$$

$$\forall i : x_i + y_i \leq L_i \qquad (4)$$

$$0 \leq x_i, y_i \leq L_i \qquad (5)$$

where L_{i} is the number of levels (discrete case) or amount of resource (continuous case). (2-5) finds the optimal efficient solution, which satisfies the fairness constraint (3). Note: (3) is a variation of Nash’s fairness axiom, known as



Imagine you are negotiating over the items shown above with another person. To you, apples are worth $3 each, bananas $2 each and pears $1 each. You don’t know your partner’s preference. Below is a hypothetical dialog between a partner and yourself. Based on your preference and the dialog provided, would you accept the offer?

Partner: Do you like apples more than bananas?

You: Yes

Partner: I like apples more than bananas as well

Partner: Do you like bananas more than pears?

You: Yes

Partner: I like Bananas more than Pears as well

Partner: I’ll offer you all the apples and keep the rest

Figure 1: An example of deceptive negotiation

Kalai's fairness. This is required for when the negotiation involves discrete levels (where an exactly equal split may not be possible). Although not obvious, this optimization problem can be reformulated and solved via mixed integer-linear programming [Nguyen and Gratch 2016].

The optimization problem handles the case where the opponent's preferences are known. We proposed a heuristic approach when the opponent's preferences are unknown. In this case, the deceiver asks preference questions that are maximally informative while offering minimally informative responses, then plays the perfect information game.

By analyzing this game, we showed that a negotiator maximally benefits by always pretending that the negotiation has a distributive (fixed-pie) structure, even if there is integrative potential. Thus, honesty is the best policy in distributive negotiations, but a malicious player can gain considerable advantage in integrative negotiations by making a "fixed-pie lie." Either way, a malicious player should simply state they have the same preferences as their opponent. For example, if the opponent says "I like X more than Y," the deceiver would match this by saying that they like X more than Y also. After building a sufficiently complete model of the opponent, the deceiver then proposes a deal that gives the opponent a disproportionate amount of the opponent's favorite issue, and keeps a disproportionate amount of the rest. This deal appears fair but benefits the deceiver.

Figure 1 illustrates this strategy for a negotiation over how to divide a basket of fruit. In this example, the negotiation has an integrative structure (the deceiver received \$3 per pears, \$2 per banana and \$1 per apple) but conveys a distributive structure. The opponent believes each side will receive \$12. In truth, the deceiver receives \$20.

An empirical study showed that more people accept this deceptive deal, and found it fairer, than if the deceiver was honest about their preferences. However, this result depended on some strong assumptions. The theoretical analysis adopted the common assumption that the value of different issues is independent, but also required the opponent to be honest. The empirical study added further constraints. Only the deceiver could ask preference questions and then offered a take-it-or-leave it ultimatum. In more realistic negotiations, parties are free to ask whatever questions they like, make tentative offers, and make counterproposals to offers they receive. Also, both parties are free to lie. Thus, it is unclear if human negotiators could pull-off this strategy in practice.

4 Model Validation

From the solution to the misrepresentation game, we can derive several predictions about how deceptive negotiators should behave. Here, we test if people actually follow these predictions. If so, the model is a reasonable approximation of human behavior and has explanatory value in practice.

First, the formal analysis indicates that malicious negotiators will benefit by claiming the negotiation is distributive (fixed-pie), regardless of the actual structure of the task. Any deviation from this policy should undermine the potential for profit. From this we can derive our first hypothesis:

> H1 (fixed-pie communication): Profits will increase to the extent that negotiators convey distributive preferences

Second, the solution to the misrepresentation problem indicates that liars only gain when the negotiation has integrative potential. If the task has distributive structure, lies only undermine the opponent's belief that the negotiation is fixed-pie. From this we can derive a two-part hypothesis:

> H2 (fixed-pie lies): Lying will enhance profit iff the negotiation has integrative potential (H2a) *and* the lies convey distributive preferences (H2b)

Third, the solution to the misrepresentation problem forces a discrepancy between words and deeds. When a negotiator is in an integrative negotiation and uses fixed pie lies, they must make an offer that creates integrative value, but keep this integrative value for themselves. This means their final offer will appear somewhat integrative. Figure 1 illustrates this. Here the liar offers their opponent more of what their opponent wants (apples), while keeping a larger amount of lower value items (where, in truth, these low-priority issues are of high value to the liar). Because of the lies, this appears to be a Nash bargaining solution (\$12 each), however there is another equivalent solution that splits the pie down the middle: each receives two apples, two bananas and two pears (also \$12 each). Although the deals are equally fair in terms of payout, the former might be seen as less consistent with fixed-pie structure. This leads to our third and hypothesis:

> H3 (evidence of lies): Liars should exhibit a discrepancy between their offers and stated preferences; truthful negotiators should not exhibit this discrepancy

Finally, a clever opponent might recognize this discrepancy and suspect lying. If people are good at detecting this sort of deception, it will be harder for malicious negotiators to succeed at the misrepresentation game. On the other hand, negotiators often assume by default that negotiations are distributive – known as a "fixed-pie bias" [Harinck et al. 2000] – which could make it especially hard to suspect fixed-pie lies. From this we derive a research question:

> RQ1 (lie detection): Do lies impact perceptions of honesty?

To test these hypotheses, we obtained a large existing corpus of dyadic face-to-face negotiations. The corpus contains negotiations with both an integrative and distributive structure. All negotiations were previously transcribed and annotated, including annotations of offers and preference statements. Thus, the corpus is ideal for testing our hypotheses.

4.1 Corpus Description

Negotiation Task: Dyads performed a simulated negotiation exercise known as "Auction Wars" (see [DeVault et al. 2015]). Each negotiator played the role of an antique dealer hoping to obtain the contents of an abandoned storage locker filled with antique items. The locker contained six antique items (three crates of LP records, two art deco lamps, and one art deco painting). Each negotiator received a private payoff matrix that explained the value of each item to them (parties did not know their opponent's preferences). Their objective was to negotiate how to divide these six items.

To elicit realistic behavior, negotiators were motivated with a financial reward. They received a number of lottery tickets corresponding to the value, for them, of the items they obtained. Higher-value items received more lottery tickets. Negotiators had 15 minutes to talk and research an agreement. If they failed to reach agreement, they received only the number of lottery tickets that they would have received for one of their highest value items. Tickets were entered into a lottery worth $100.

Negotiators engaged in one of two variations of this task. In the integrative task, payoffs allowed win-win solutions (one side wanted the records the most, whereas the other side most-wanted the lamps). In the distributive variant of the task, negotiators had to divide a fixed pie (both sides wanted the records the most). In both variants, the painting is of little or no value (see Table 1).

Participants: The corpus consists of 226 participants (113 same-sex dyads; 66% male) that were recruited from craigslist.com. In addition to the lottery, participants were compensated $30 for completing the study. Of the 226 participants, 8 participants did not comply the procedure and 28 failed to correctly report their own preferences at the end of the study. As we wish to focus on intention misrepresentation, these participants were excluded from our analysis, resulting in 190 participants. Participants also provided subjective ratings on how honest their opponent seemed.

Annotations: The corpus was manually annotated. All negotiations were segmented, transcribed and semantically annotated by trained coders using ELAN [Wittenburg et al. 2006]. Here, we describe only the subset of annotations relevant to our purposes. An utterance was noted as a division-of-items (DIV) if it suggested a full or partial partition of the items (e.g., "How about if I take two records and you take both lamps?"). An utterance was noted as a preference statement (PREF) if it asserted some preference over the items (e.g., "I like lamps the most" or "I like the painting more than the records."). On average, negotiations contained 5.8 DIVs and 3.8 PREFs.

Of the 190 participants, 34 could not be analyzed for lying because the participants failed to make any PREFs (they simply exchanged offers without explicitly exchanging preference information), and 10 participants (5 dyads) failed to reach an agreement within the 15-minute time limit. We exclude these, retaining 146 participants for the analysis below (73 integrative and 73 distributive).

4.3 Measures

Hypothesis H1 requires a measure of winnings and a notion of what preferences are communicated. Hypotheses H2 requires measures of the number and type of lies elicited. H3 requires a measure of the "distance" between preference statements and offers. Finally, the research question requires subjective statements of honesty, which were already provided as part of the corpus.

Task	**Partner**	**Records**	**Lamps**	**Painting**
Distributive	A	High (30)	Medium (15)	Low (5)
	B	High (30)	Medium (15)	Low (0)
Integrative	A	High (20)	Medium (10)	Low (5)
	B	Medium (10)	High (30)	Low (0)

Table 1. Payoffs for different issues across task and partner.

Winnings: A negotiators objective success can be measured in terms of the number of lottery tickets they earn. Connecting this to the misrepresentation game, we can see from Equation (1) and Table 1, the value of a negotiated deal is the sum from $i=1..3$ of $a_i\ .\ x_i$ where a_i represents the importance of an issue (i.e., the number of lottery tickets for the i^{th} issue) and x_i is the agreed level on that issue (number of items obtained). As there are 3 records, 2 lamps and 1 painting, the values of c_i in Equation (3) are 3, 2, and 1 respectively.

Lies: As a measure of explicit lies, we manually annotated all preference statements for their veracity. For each of the 342 preference statements, we recorded it as a lie if it was inconsistent with the participant's private preference. For example, if a participant stated they liked lamps most, but in truth they like records more than lamps, this would be recorded as a lie. A lie was classified as a "fixed-pie lie" if it matched the opponent's preferences. For example, if 1) a participant liked records most, and 2) their opponent liked lamps most, and 3) the participant stated they liked lamps most, this statement was noted as a fixed-pie lie.

These annotation do not capture all possible lies, but only misrepresentation of relative imporance of preferences. Negotiators could lie about other things, such as their payout if the negotiation fails (we return to this at the conclusion).

Opponent model (from words): The previous measure indexes lies of commission (i.e., explicit lies) but there are many ways to convey a false impression of one's preferences. For example, if a negotiator repeatedly talks about how much they like the lamps, their opponent might assume the lamps are their most important item. AI research in opponent-modeling attempts to infer an opponent's preferences by analyzing their behavior [Baarslag et al. 2015]. Indeed, prior opponent-modeling research has shown that simply counting the number of times issues are mentioned can provide a surprisingly accurate estimate of a human opponent's preferences [Nazari et al. 2015]. Using these models, we can objectively measure the type of preferences people are communicating (H1) as well as the differences between what people communicate between words and deeds (H3).

To objectively measure the preferences people are communicating with their words (including both explicit and implicit lies), we adapt [Nazari et al. 2015]'s *issue-sentiment* opponent-modeling heuristic. This heuristic simply counts the number of positive and negative PREF statements an opponent makes about each issue. The relative counts across issues correspond to their estimated weight (see [Nazari et al. 2015] for precise definitions).

Using this opponent model, we can derive a measure of whether a person's verbal communication is communicating a distributive structure. To do so, we compare the distance between the ranking discovered by the issue-sentiment heuristic and the ranking the party should have communicated if they wanted to convey a fixed-pie ranking. For example, if one party talks most about lamps and least about the painting, this would imply the ranking: Lamps > Records > Painting. If their opponent has, in truth, this same ranking, then the distance between these two ranks would be zero. In other words, a small distance implies fixed-pie communication.

Specifically, we use a standard distance metric called rank distance to measure the distance between two rankings. This compares the utility of all possible deals in the outcome space (Ω), given the first model (u_{op}) and the second one (u'_{op}), and calculates the average number of conflicts in how deals are ranked using each of the two models' utility functions:

$$d_r(u_{op}, u'_{op}) = \frac{1}{|\Omega|^2} \sum_{\omega \in \Omega, \omega' \in \Omega} c_{<u,<u'}(\omega, \omega') \qquad (2)$$

From this, we define a measure of the "rank-distance from fixed-pie" (or RDFP-Words) that captures the distance between the weights estimated by issue-sentiment heuristic and the weights a negotiator should have communicated based on the fixed-pie lie tactic. Low values of RDFP-Words imply the negotiator is claiming distributive preferences.

Opponent model (from deeds): The issue-sentiment heuristic gets at what preferences people communicate with their words, but people also communicate their preferences indirectly through their pattern of offers. We adapt [Nazari et al. 2015]'s *issue-ratio* heuristic to infer a negotiator's preferences from the offers (DIVs) they make throughout the negotiation. Intuitively, the issue-ratio assumes that people try to be fair in their offers and any deviation from their perceived fair point could be a hint for what they most value. The heuristic estimates the weight of issue *i* by dividing the average level kept for *i* by the average level kept across all issues.

From this we derive a measure of the "rank-distance from fixed-pie" (or RDFP-Deeds) to see how far an opponent's offers diverge from what they should offer in a distributive (fixed-pie) negotiation.

Honesty Perceptions: To address the research question, we need to measure the perceived honesty a negotiator has of their opponent. For this, we use the self-reported honesty ratings include in the corpus (a 7-point Likert item).

4.2 Results

Lies: The misrepresentation game can only succeed if most negotiators are (more-or-less) honest, but we can only test the model's predictions if at least some people lie. Fortunately, both properties hold in this corpus. Most participants were honest, but lying was evident. Only 30 of the 146 participants lied (21%). Of these, 15 performed the distributive negotiation (21%) and 15 performed the integrative negotiation (21%). Only one dyad had two liars. Of 342 preference statements, 46 were lies (13%). Similar rates of lying were found in each negotiation type: 21 of 158 in distributive (13%) and 26 of 184 in integrative (14%).

H1: The first hypothesis states that profits will increase to the extent that negotiators convey distributive preferences (ignoring if they are lying or honest). To test this, we examined the overall sentiment people communicated (RDFP-Words) which captures both explicit and implicit preference statements. We found a significant negative correlation between RDFP-Words and the number of lottery tickets earned across the entire corpus (r(146) = -0.247, p = 0.003). In other words, the closer their verbally-communicated preference was to a fixed pie, the better they did. To see how this effect interacts with the structure of the negotiation, we performed a moderated regression, predicting earnings using both RDFP-Words and structure (dummy coded: 0 = integrative , 1 = distributive) in a first step, followed by the RDFP-Words × structure interaction term in a second step. This analyses showed a significant main effect of RDFP-Words, as predicted by the correlation analysis (β=–.20, t(143)=–2.43, p=0.02), as well as significant main effect of structure (β = 0.16, t(143)=1.96, p=0.05). In other words, participants win more money in distributive negotiations. The interaction of RDFP-Words × structure, however, did not reach significance (β=0.17, t(143)=1.56, p=0.12), suggesting that RDFP-words predicts profit regardless of the negotiation structure. This supports our first hypothesis.

H2: The second hypothesis asserts that lying will only enhance profit if (H2a) the negotiation has integrative potential and (H2b) if the lies are in a fixed-pie direction. For H2a, we conducted two separate t-tests for integrative and distributive negotiations. In the integrative case, liars earned significantly more lottery value (M=63.33, SD=9.19) than non-liars (M=55.12, SD=13.91); t(71)=2.159, p=0.02. In the distributive case, however, the difference in earnings between liars (M=64.33, SD=18.60) and non-liars (M=63.19, SD=15.57) as insignificant; t(71)=0.24, p=0.30. This result supports hypothesis H2a.

Next, we examined the type of lies told in the integrative negotiations. Of the lies 25 lies elicited in the integrative task, virtual all (88%) were fixed-pie lies. (Note that there cannot be fixed-pie lies in the distribution task. As the task, in truth, has a fixed-pie structure, any lie would necessarily claim something other than a fixed-pie.) Together, these two findings confirm hypothesis H2.

As second way to examine H2 is to look at the explicit and implicit communication made by negotiators. For this, we examined the overall sentiment people communicated (RDFP-Words) separately for integrative and distributive negotiations. We found a significant correlation between fixed-pie communications and profits for integrative negotiations (r(73) = -0.365, p = 0.001), but no correlation in the distributive negotiations (r(73) = 0.045, p=0.704). In other words, the closer their verbally-communicated preference was to a fixed pie, the better they did, but only or integrative negotiations. Again, this supports hypothesis H2.

H3: Our third hypothesis asserts that liars give off indications of lying. Specifically, deceivers should be more distributive in their words but more integrative in their offers. To test this, we divided participants into liars and non-liars and

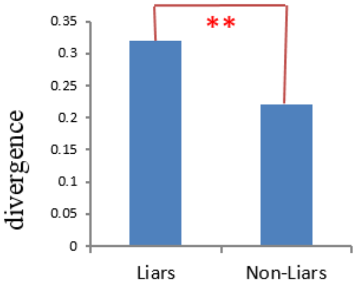


Figure 2. Divergence between the issue-ratio and isssue-sentiment models. Liars: (M = 0.32, SD = 0.19), Non-Liars: (M =0.22, SD = 0.12), t(144) = 3.54, p = 0.001

calculated the rank distance between sentiment-ratio and issue-ratio to see if there is a significant difference between the liar and non-liar groups. A t-test shows that indeed liars have bigger difference between their words and deeds than the non-liars group (see Figure 2). This supports hypothesis H3.

Research question: Support for H3 suggests it is theoretically possible to detect malicious actors. But do participants recognize this in practice? Our final research questions ask if participants found liars to be any less honest than truth-tellers. To test this, we again divided participants into liars and non-liars and performed t-tests to compare perceived honesty of their opponent (a single-item 7-pt licker scale). Liars were considered significantly less honest in the distributive condition: liars (M = 5.33, SD = 1.67) vs. non-liars (M = 6.24, SD = 0.93), t (67) = -2.75, p=0.008. In the integrative condition, liars were actually considered more honest than honest participants – liars (M = 6.21, SD = 1.18) vs. Non-Liars (M = 5.89, SD = 1.42), t(65) = 0.79, p=0.432 – but this difference was not significant. Recall that in the integrative condition, lies conform with the common preconception that negotiations are fixed-pie (i.e., the "fixed-pie bias" [Harinck et al. 2000]). But in the distribution condition (which truthfully is a fixed pie), any lies are in conflict with the fixed-pie bias. This suggests that fixed-pie lies are especially difficult for negotiators to detect.

5 Discussion

The results provide strong evidence that the misrepresentation game models the behavior of human negotiators, even in situations that don't obey the strict assumptions of the model. If negotiators use the fixed-pie lie tactic, they won. Interestingly, they were seen as equally honest (and were actually rated as more honest that truthful negotiators, though this difference was not significant). We also confirmed that lying about one's preferences only confers a benefit if the negotiation has integrative potential. If the negotiation has a distributive structure, lying failed to confer any monetary benefit, and actually harmed their reputation (as liars were perceived as significantly less honest in the distributive condition).

These results suggest that self-interested human negotiators or even software agents can exploit a common cognitive bias (the "fixed-pie bias" [Harinck et al. 2000]) and the human proclivity for fairness, to earn disproportionate rewards while seeming honest and fair. The better they conform to the fixed-pie lie, the better they will do.

Fortunately, there is a silver-lining for honest-minded negotiators. Automated analysis was able to find evidence of lying by looking for discrepancies between the preferences communicated through words vs. deeds. Unfortunately, human negotiators completely failed to notice this discrepancy. But perhaps through training they could learn to defeat this tactic (e.g., see [Reilly 2009]).

Finally, it is worth noting that not all liars were consistent with the model. A small number of lies in the integrative negotiations (12%) and all the lies in the distributive negotiations were inconsistent with the model. But liars that failed to follow the model also failed to better their profits. This provides more evidence that the solution to the misrepresentation game serves as a good prescriptive model for malicious actors, and this suggests that their effectiveness could (unfortunately) improve with training.

There are several limitations to the current work. First, the "fixed-pie-lie" solution to the misrepresentation problem proposed depends on an assumption of issue-independence. In other words, the value obtained for one issue does not depend on the value for other issues. Auction Wars (the negotiation task we examined in this paper) obeys this assumption. When negotiations contain interdependent issues, the misrepresentation game might yield other solutions, and dramatically increase the computational cost of finding them. We need to test the value of these predictions in these situations as well.

A second limitation is that the misrepresentation game focuses on a specific type of lie (misrepresenting the ranking of one's preferences). But other lies are possible. Negotiators can lie about their best alternative to the current negotiation. For example, when buying a car, a prospective buyer might exaggerate the deal they received from another car dealership. Negotiators can also make false promises about implementing negotiated agreements. For example, an employer might promise flexible time but take back that promise once an employee joins. It remains to be seen which of these lies would offer the greatest rewards to malicious negotiators, and how to defend against them.

Finally, our findings show that a system could theoretically identify attempts at deception but we have yet to show this convincingly. Further, there may be other tactics to inoculating oneself from exploitation. For example, at least in the Auction Wars task, a fair-minded negotiator could avoid exploitation by insisting on splitting the items down the middle. Thus, a fuller consideration of defensive tactics is needed. This can include deception detecting and defensive actions, but also the design of mechanisms that eliminate the incentive to misrepresent (e.g., see [Procaccia 2013]).

The current paper focused on strategic misrepresentation in negotiation but lying arises in many contexts. For example, Naamani-Dery explore similar issues that can arise in the context of elections [Naamani-Dery et al. 2015]. More broadly, deception arises across a variety of real-world social dilemmas [Burton-Chellew and West 2012]. Linking models from across different contexts could benefit systems that aim to broadly understand human social behavior.

In summary, we showed that the misrepresentation game does have explanatory value in modeling deceptive negotiators. Even in a realistic negotiation with far fewer restrictive assumptions, human behavior was consistent with the model. Not all malicious actors lied in the optimal way, but when they did, they reaped considerable reward.

Acknowledgments

This work was supported by the NSF (BCS-1419621) and the U.S. Army. Any opinion, content or information presented does not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

[Baarslag et al., 2015] Baarslag, T., Hendrikx, M., Hindriks, K. and Jonker, C. 2015. Learning about the opponent in automated bilateral negotiation: a comprehensive survey of opponent modeling techniques. *Autonomous Agents and Multi-Agent Systems*: 1-50.

[Burton-Chellew and West, 2012] Burton-Chellew, M. N. and West, S. A. 2012. Correlates of cooperation in a one-shot high-stakes televised prisoners' dilemma. *PloS one* 7(4): e33344.

[Colman, 2003] Colman, A. M. 2003. Cooperation, psychological game theory, and limitations of rationality in social interaction. *Behavioral and brain sciences* 26(02): 139-153.

[DeVault and Gratch, 2015] DeVault, D., Mell, J. and Gratch, J. 2015. Toward Natural Turn-Taking in a Virtual Human Negotiation Agent. *AAAI Spring Symposium on Turn-taking and Coordination in Human-Machine Interaction*, Stanford, CA, AAAI Press.

[Fatima, et al., 2014]Fatima, S. S., Wooldridge, M. and Jennings, N. R. 2004. An agenda-based framework for multi-issue negotiation. *Artificial Intelligence* 152(1): 1-45.

[Fehr and Schmidt, 1999] Fehr, E. and Schmidt, K. M. 1999. A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics* 114(3): 817-868.

[Gratct, et al, 2016] Gratch, J., Nazari, Z. and Johnson, E. 2016. *The Misrepresentation Game:* How to win at negotiation while seeming like a nice guy. *International Conference on Autonomous Agents and Multiagent Systems*, Singapore.

[Harinck, et al., 2000] Harinck, F., De Dreu, C. K. W. and Van Vianen, A. E. M. 2000. The Impact of Conflict Issues on Fixed-Pie Perceptions, Problem Solving, and Integrative Outcomes in Negotiation. *Organizational Behavior and Human Decision Processes* 81(2): 329-358.

[Hindricks and Tykhonov, 2008] Hindriks, K. and Tykhonov, D. 2008. Opponent modelling in automated multi-issue negotiation using bayesian learning. *Proceedings of AAMAS*.

[Hindricks and Jonker, 2008] Hindriks, K. V. and Jonker, C. M. 2008. Creating human-machine synergy in negotiation support systems: Towards the pocket negotiator. *Proceedings of the 1st International Working Conference on Human Factors and Computational Models in Negotiation*, ACM.

[Jennings, et al., 2001] Jennings, N. R., Faratin, P., Lomuscio, A. R., Parsons, S., Sierra, C. and Wooldridge, M. 2001. Automated Negotiation: Prospects, Methods and Challenges. *International Journal of Group Decision and Negotiation* 10(2).

[Kraus et al., 2008] Kraus, S., Hoz-Weiss, P., Wilkenfeld, J., Andersen, D. and Pate, A. 2008. Resolving crises through automated bilateral negotiations. *Artificial Intelligence* 172(1): 1-18.

[Naamani-Dery, et al., 2015] Naamani-Dery, L., Obraztsova, S., Rabinovich, Z. and Kalech, M. 2015. *Lie on the fly: iterative voting center with manipulative voters*. *Proceedings AAAI*, AAAI Press.

[Nash Jr, 1950] Nash Jr, J. F. 1950. The bargaining problem. *Econometrica: Journal of the Econometric Society*: 155-162.

[Navari et al., 2015] Nazari, Z., Lucas, G. and Gratch, J. 2015. Opponent Modeling for Virtual Human Negotiators. *15th Conference on Intelligent Virtual Agents*, Delft, Springer.

[Nguyen and Gratch, 2016] Nguyen, T. and Gratch, J. 2016. Misrepresentation Negotiation Games, Playa Vista, CA, University of Southern California Institute for Creative Technologies.

[O'Connor and Carnevale, 1997] O'Connor, K. and Carnevale, P. 1997. A Nasty but Effective Negotiation Strategy: Misrepresentation of a Common-Value Issue. *Personality and Social Psychology Bulletin* 23(5): 504-515.

[Paulhus and Willimas, 2002] Paulhus, D. L. and Williams, K. M. 2002. The dark triad of personality: Narcissism, Machiavellianism, and psychopathy. *Journal of research in personality* 36(6): 556-563.

[Procaccia, 2013] Procaccia, A. D. 2013. Cake cutting: not just child's play. *Communications of the ACM* 56(7): 78-87.

[Reilly, 2009] Reilly, P. R. 2009. Was Machiavelli right? Lying in negotiation and the art of defensive self-help. *Ohio State Journal on Dispute Resolution* 24(3): 09-05.

[Rosenfeld et al., 2014] Rosenfeld, A., Zuckerman, I., Segal-Halevi, E., Drein, O. and Kraus, S. 2014. NegoChat: a chat-based negotiation agent. *Proceedings of AAMAS*, Paris, France.

[Schweitzer, et al., 2002] Schweitzer, M. E., Brodt, S. E. and Croson, R. T. A. 2002. Seeing and believing: visual access and the strategic use of deception. *International Journal of Conflict Management* 13(3): 258-375.

[Thompson, 1991] Thompson, L. L. 1991. Information exchange in negotiation. *Journal of Exp. Social Psychology* 27(2): 161-179.

[Traum et al., 2008] Traum, D., Gratch, J., Marsella, S., Lee, J. and Hartholt, A. 2008. Multi-party, Multi-issue, Multi-strategy Negotiation for Multi-modal Virtual Agents. *8th International Conference on Intelligent Virtual Agents*, Tokyo, Japan, Springer.

[Wittenburg, et al., 2006] Wittenburg, P., Brugman, H., Russel, A., Klassmann, A. and Sloetjes, H. 2006. Elan: a professional framework for multimodality research. *Proceedings of LREC*.