Journal of Artificial Intelligence Research 68 (2020) 633-660 Submitted 01/2020; published 07/2020

The Effects of Experience on Deception in Human-Agent Negotiation

Johnathan Mell MELL@ICT.USC.EDU
Gale M. Lucas LUCAS@ICT.USC.EDU
Sharon Mozgai SMOZGAI@POST.HARVARD.EDU
Jonathan Gratch GRATCH@ICT.USC.EDU
Institute for Creative Technologies, University of Southern California
Los Angeles, CA, 90066

Abstract

Negotiation is the complex social process by which multiple parties come to mutual agreement over a series of issues. As such, it has proven to be a key challenge problem for designing adequately social AIs that can effectively navigate this space. Artificial AI agents that are capable of negotiating must be capable of realizing policies and strategies that govern offer acceptances, offer generation, preference elicitation, and more. But the next generation of agents must also adapt to reflect their users' experiences.

The best *human* negotiators tend to have honed their craft through hours of practice and experience. But, not all negotiators agree on which strategic tactics to use, and endorsement of deceptive tactics in particular is a controversial topic for many negotiators. We examine the ways in which deceptive tactics are used and endorsed in non-repeated human negotiation and show that prior experience plays a key role in governing what tactics are seen as acceptable or useful in negotiation. Previous work has indicated that people that negotiate through artificial agent representatives may be more inclined to fairness than those people that negotiate directly. We present a series of three user studies that challenge this initial assumption and expand on this picture by examining the role of past experience.

This work constructs a new scale for measuring endorsement of manipulative negotiation tactics and introduces its use to artificial intelligence research. It continues by presenting the results of a series of three studies that examine how negotiating experience can change what negotiation tactics and strategies human endorse. Study #1 looks at human endorsement of deceptive techniques based on prior negotiating experience as well as representative effects. Study #2 further characterizes the negativity of prior experience in relation to endorsement of deceptive techniques. Finally, in Study #3, we show that the lessons learned from the empirical observations in Study #1 and #2 can in fact be induced—by designing agents that provide a specific type of negative experience, human endorsement of deception can be predictably manipulated.

1. Introduction

Increasingly, humans interact with the world around them, and with each other, through artificially intelligent intermediaries. Agents coordinate our call centers, suggest new friends and connections, supervise our economic markets, sort our mail, and are beginning to drive our cars. But this brave new world of socially-aware agents raises several important questions. What social and ethical norms should govern these interactions? Can these be quantified and represented within intelligent systems? What control should people have over these norms? And, ultimately, does this change the nature of how people treat each other? If it does, we have much to learn experimentally before we can design socially-aware agents that are effective.

This article explores these questions within the context of negotiation. Negotiations are interactions between two or more parties – each with its own aims, needs, and viewpoints – seeking to find an acceptable common ground. This sort of give-and-take is ubiquitous in human social interactions but is increasingly the focus of artificial systems that can negotiate on the behalf of human users (Baarslag et al., 2017). Many of these human-AI interaction tasks can be broadly characterized as negotiations: exchanges between two (or more) parties in which agreement is reached over a set of issues. Artificially-intelligent agents have shown success at these kinds of tasks, automating online bidding (Anthony & Jennings, 2003), providing human-driven ethical decision-making to autonomous vehicles (Bonnefon et al., 2016), and even negotiating the time of appointments for their users (Metz, 2018). The time at which automated agents will largely be able to represent us is nigh (Baarslag et al., 2017).

From the standpoint of this article, negotiations are interesting in that they evoke strong differences of opinion on what are acceptable social and ethical norms. Some professional negotiators and ethicists argue that unfairness, deception and emotional manipulation are perfectly acceptable and expected behavior (Levine & Schwitzer, 2014; White, 1980). Thus, we examine which norms could govern automated negotiators, the extent to which users would wish to customize behavior of AI systems that might be counter-normative/unethical, and how this might shape human interactions as people begin to negotiate with each other through AI systems.

In a recent provocative paper in the Academy of Sciences, de Melo and colleagues examined the norm of fairness and showed that people treated each other more fairly when they interacted through AI systems (de Melo et al., 2018). Using several domains, including negotiation, they found that when people "programmed" the behavior of an agent to act on their behalf, the agent treated others more fairly than they would have themselves. This suggests that people, if allowed to customize their agent's propensity to lie and manipulate, would choose to create agents that are paragons of moral behavior. Yet this seems hard to reconcile with evidence that people often instruct human intermediaries (e.g., lawyers) to act less than ethically on their behalf (Chugh et al., 2005). This article seeks to lend insight into this apparent contradiction.

We make several contributions. First (Section 4), we introduce a way to measure the extent to which people are comfortable with deceptive tactics (e.g., lying, withholding information, and emotional manipulation). We present the Agent Negotiation Tactics Inventory (ANTI). This inventory both serves as a measure of what tactics people are willing to use or endorse, but also a "programming language" of sorts by which people can constrain the behavior of agents that negotiate on their behalf.

Next, in a series of empirical studies, we examine the extent to which people are willing to endorse manipulative tactics, if these endorsements differ when they negotiate directly with others or through

automated negotiations, and how negotiation experience might alter these same endorsements. We find (Section 5.1) that novices are more comfortable with deception when agents engage in such manipulation on their behalf, but that experts are comfortable with morally questionable tactics, regardless of whether they interact directly or through technology. In a second study (Section 5.2), we look more closely at the nature of people's experiences in negotiation. In an empirical study, we find evidence that negotiators are more willing to engage in deceptive tactics when their experience with negotiation is negative.

Finally (Section 5.3), we show that even short term experience can significantly shape the ethical frame people adopt for agents that act on their behalf. Participants are given the opportunity to program the "ethics" of an agent that will negotiation with another person on their behalf for an opportunity to earn real money. They first program their agent using the items from the ANTI. Before launching their agent, they have the opportunity to first negotiate with another person's agent and then change their own programming based on this concrete experience. Unbeknownst to participants, we manipulated the negativity of the negotiation experience (using two different operationalizations). We find that this short experience (i.e., a single negotiation) significantly impacts the ethical choices they adopt for their own agent. People who had a negative experience (in particular, interacting with a tough negotiator) were much more willing to tell their agent to lie and engage in emotional manipulation.

Together, these results indicate several key points. First, endorsement of ethical techniques varies considerably across individuals, but there are general trends in how much people are willing to endorse deceptive behaviors. Second, while the act of programming an agent representative rather than negotiating directly does have a key effect on endorsement of deception, this relationship is entwined with experience, and specifically negativity of experience. Finally, these effects are mutable, and may be "nudged" by strategically-curated agent behaviors—leading to implications for the design of agents that may try to understand or exploit this fact in the future. We successfully demonstrate a number of agents that vary according to our proposed subscales within the ANTI; by creating these socially-varied agents, we craft a toolbox of effective agents that can better be applied to the individual.

2. Related Work

2.1. Human-Agent Negotiation and Representation

Negotiation is a broad field that entails a great deal of social intelligence on behalf of its participants. Indeed, exploration of the myriad social effects that make good negotiators effective is well supported by decades of research into human-human negotiation techniques from the business and psychological literatures (Kelley, 1966; Patton, 2005; Raiffa, 1982), Within computer science, there have been numerous attempts to formalize negotiation processes, and to characterize both the outcomes and procedures of negotiation. A number of unifying protocols have been proposed (Fatima et al., 2006; Zlotkin & Rosenschein, 1996) for describing negotiation between agents. And while social welfare is generally seen as the province of human-human or human-agent negotiation, it has been examined even in agent-agent contexts. Specifically, social welfare is considered an important outcome for negotiation (Endriss et al., 2006), and indeed, agents have been shown to be able to come to social agreement in certain tasks, like the Ultimatum game (De Jong et al., 2008). But while these optimizations show progress in designing "ideal" agents, they do not directly address the question of how humans actually act (Wright & Leyton-Brown, 2014). Furthermore, there is an increasing desire to have agents that work in tandem with humans (Scerri et al., 2002) or (going further) that are directly programmed with (explainable)

behaviors by humans (Zanzotto, 2019). As such, it is important to understand what choices humans are likely to make in various situations. Otherwise, the agents designed to interact with humans may not make reasonable or realistic decisions when actually deployed as functional systems.

In behavioral literatures, previous work has examined a specific type of social interaction: where one person acts on behalf of a human client as their representative. Often, when people represent others, they are directed to follow specific policies and norms. This is a readily observable phenomenon—many people see the value in hiring a lawyer, a real estate broker, or other representative to convey their interests. These instructions may range from the specific ("I won't pay more than $5000 up front!") to the general ("I'm buying this from a family friend, so it's important everyone walks away happy!"). And indeed, there is considerable debate on which policies are ethical to follow if instructed by one's client (White, 1980).

There is a curious effect of this kind of indirect "middleman" interaction: the instructions that principals provide may not be the same ones they would themselves follow if they were negotiating directly. There is some evidence that individuals instruct their representatives to perform more fairly than they themselves would due to reputation (Ramchurn et al., 2003) or temporal effects (Pronin et al., 2008)—due to a desire to be perceived by others as good or fair—but these effects persist even into anonymized scenarios (de Melo et al., 2016). Some theories maintain that, when considering what instructions to provide to one's representative, principals engage in higher level thinking about fairness and equity and other broad social goals than they might otherwise do in the heat of the moment (de Melo et al., 2018; Giacomantonio et al., 2010). Indeed, this may be the goal of some people who value indirect interactions—allowing "cooler heads to prevail" can often have direct benefits for all parties. Of course, some other results indicate the opposite effect—increasing social distance through the use of a representative of any type could reduce cooperation and fairness (Trope & Liberman, 2010). The picture of people's ethical preferences around representatives is thus somewhat incomplete—and may depend largely on context and experience. Some prior work goes so far as claim that the process of informing a representative by having users programming an artificial agent may lead to positive social change, since people will be fairer to each other (de Melo et al., 2018). These prior results lead to the idea that by framing the task of endorsing strategies as programming an artificial agent, people will be less likely to endorse manipulative techniques than if the task were framed as stating personal preferences.

This work therefore scrutinizes the fuzzy relationship between a human principal[1] and his/her computerized, agent representative. We examine how human principals' opinions on negotiation tactics are formed from their initial negotiation experiences, and then how they unfold over time. This work thus builds on the developing, but as of yet incomplete picture of social norms, experience, and representative effects within human and human-agent negotiation. This picture is necessary for the development of effective socially-aware agents.

Specifically, in this article, we ask human participants about their willingness to endorse a number of negotiation tactics according to their own interests and social norm opinions. We examine how the amount and type of experience affects these preferences. Then, we manipulate experience by exposing subjects to an online negotiation with one of four different social agents, each utilizing a different

[1] From the legal lexicon, the person hiring a representative is called the "principal". Throughout this paper we will refer to the two human parties as principal and representative, in order to avoid confusion with the common term "agent", which will be used to solely refer to computerized, artificial agents serving as representatives.

strategy. Finally, principals are asked again to report their endorsement of tactics, and the resulting changes are reported and analyzed in the context of the agent strategy the participants encountered.

2.2. Multi-Issue Bargaining

One classical option for investigating human-agent negotiation takes the form of the multi-issue bargaining task, which is considered a de facto standard problem for research into social cognition and interpersonal skill training (Van Kleef et al., 2004). In the multi-issue bargaining task, two participants work to determine how to split varying issues, each participant having hidden values unknown to the other side. The task may involve distinct phases, where first information about preferences is exchanged, and then a series of offers are made. The task is also often characterized by time pressure, which is often modeled as a decaying utility function. Even with a small number of issues, the task can quickly become a challenge for agents to simulate, especially those which aim to act as partners for humans in such a negotiation in real time, and numerous works attempt to address the multi-issue bargaining task (Fatima et al., 2007; Peled et al., 2011; Robu et al., 2005). While this makes the multi-issue bargaining task a difficult challenge computationally, adding a human actor complicates issues even further, since humans often behave "irrationally" in game theoretic contexts.

2.3. Agent Programming and Strategy

There have been a number of related studies that attempt to examine the specific act of instructing a representative. In the case of an automated agent, this instruction phase sometimes takes the form of high-level programming of the agent. What constitutes "programming" varies considerably across related work—ranging from fully specifying responses to the entire domain of potential offers to more general instructions (Chalamis et al., 2013; Elmalech & Sarne, 2013; Elmalech et al., 2014; Lin et al., 2010). Research indicates that the act of programming an agent can lead to differing results than does merely deciding one's own preferences. Moreover, people do treat agents differently than they do humans, and this varies with the realism of the agent (e.g., how much affect or intelligence it demonstrates)—these and other framing effects regarding human vs. agents have been well established as far back as (Reeves & Nass, 1997). In this work, we hold the opposing party constant: participants are always competing against an opponent's agent. But we examine the framing effect of programming an agent versus merely providing your own preferences (without an intermediary representative) in Studies #1 and #2. In Study #3, this representative effect is held constant across conditions, while the effect of induced experience is examined by changing the characteristics of the opposing agent.

Our first experimental study (see Section 5.1) provided answers to these questions. We found that some deceptive techniques were endorsed less readily than more "positive" techniques, and furthermore demonstrated the link between past experience and endorsement of deception. However, the framing effect of programming an agent was significant in the *opposite* direction of prior work.

These results led us to suspect that this picture of purely static endorsements may be incomplete. Although framing and past experience may lead to certain initial goals and values for baseline endorsements, prior work does not account for the role of short-term, highly-valanced[2] experiences. Much as short-term mood is differentiated from long-term personality traits (Meyer & Shack, 1989), there is no reason a priori to expect that endorsement of techniques cannot be influenced by recent experience,

[2] "Highly-valanced" here refers to very strong positive or negative emotions. We interpret the results of this work to suggest that these types of experiences are significant in altering people's levels of endorsement.

especially among those who are not already highly experienced. Since agents are capable of numerous roles within a negotiation (including creating and molding experience itself), this has design implications for social agents. We note additionally that there is a developing literature on *repeated* negotiations, and that deceptive endorsement may change when future, repeated tasks are likely—this is not, however, the focus of the work in this manuscript.

3. Method

In this work, we focus on one constrained, but well-supported area of human negotiating behavior: manipulation strategies. Not all humans act in similar ways. While the best human negotiators may indeed choose tactics from a "playbook", levels of experience in negotiation vastly differ across negotiators. Accordingly, which strategies are used vary as well. Yet, even when informed of a number of different tactics that could be used, humans differ further on which tactics they *choose* to use. It should not be unexpected that this may also differ due to levels of prior experience, yet this question has not yet been fully explored. What norms govern which strategies are seen as acceptable? Are certain types of strategies classifiable within groups? Can we predict which strategies people may use?

As stated above, previous work has indicated that people that negotiate through artificial agent representatives may be more inclined to fairness than those people that negotiate directly. We present a series of three user studies that challenge this initial assumption and expand on this picture by examining the role of past experience. We also go even further by directly modifying experience through exposure to diverse types of agents that demonstrate different sets of manipulative behaviors.

In Study #1, we present the results of a user study in which human participants are asked about their endorsement of various negotiating tactics as well as their previous experience with negotiating. We manipulated the description of the task, such that some participants were told that they were providing information about what tactics they endorsed directly, and the remaining participants were told that the information provided would be guidelines used to program an artificial agent on their behalf. We examined this agent representative/human framing effect, as well as the effects of prior participants' experience on their endorsement of deceptive techniques. The results indicate that not only does prior experience predict increased endorsement of deceptive techniques, but also that framing the interaction as a representative programming task leads to a similar increase in endorsement of deception (especially among those who are not highly experienced). This provides evidence that against the opposite view held from prior work that this kind of a task framing increases feelings of fairness.

We suspected that recent, relatively short negotiation experiences can considerably alter what manipulative techniques humans are willing to endorse. We first aimed to confirm our suspicion that specifically *negative* experiences will have the most effect in shifting endorsement. In Study #2, we distinguish between positive and negative prior experience. Participants are subject to the same manipulations and questions as Study #1. However, participants are additionally asked to characterize their past negotiating experience, and this measure of experiential negativity is used to add distinctions to the results from Study #1. We found that there was a main effect such that experience in general led to increases in reported positivity of past experience. But, controlling for this effect, negativity of experience did indeed predict greater endorsement of deceptive techniques, supporting our initial suspicions. This mechanism could either be due to exposing humans to techniques of which they were previously unaware, or by causing humans to retaliate by punishing poor behavior, or by creating a negative or pessimistic view of negotiation in general and how one should act therein. The effect of negativity

breeding similar counter-reactions is supported by the highly effective nature of "tit-for-tat" strategies in human negotiation (Kreps et al., 1982).

We speculate in Study #2 that more pessimistic views of negotiation might drive increased use and endorsement of deceptive tactics. The goal of Study #3 is to directly control experience negotiating, in order to directly determine how this experience affects the endorsement of more deceptive tactics. In particular, "negative experience" can take many forms. Previous human-agent interactions have differentiated between negative "words" and negative "deeds" (Mell & Gratch, 2017). In this final experiment (Study #3), we distinguish first between negativity of experience with agent attitude (an agent's "word"), and negativity of experience with agent hard bargaining (an agent's "deeds"). In so doing, we examine the effects of interacting with four different types of automated agents, each with a unique strategy, and how this subsequently changes which strategies a human negotiator might later endorse.

In the study, which was conducted on an online negotiation platform, four different types of automated agents negotiate with humans over the course of a 10-minute interaction. The agents differ in a 2x2 design according to agent strategy (tough deeds vs. fair deeds) and agent attitude (nice words vs. nasty words). These results show that in this multi-issue bargaining task, humans that interacted with a tough agent were more willing to endorse deceptive techniques when instructing their own representative. These kinds of techniques were endorsed even though the agent the human encountered did not use deception as part of its strategy. In contrast to some previous work, there was not a significant effect of agent attitude. These results indicate the power of allowing people to program agents that follow their instructions, but also indicate that these social norms and tactic endorsements may be mutable in the presence of a recent, impactful negotiation experience. By exposing human negotiators to tough, automated agents, we are able to shift the participant's willingness to deceive others and utilize "hard-ball" negotiation techniques. In short, which manipulative techniques people decide to endorse is dependent upon their experience—and that experience is highly controllable in the short-term.

4. The Agent Negotiation Tactics Inventory

How humans say they will make decisions and how they actually make decisions are rarely aligned. This is especially true in negotiation, in which negotiators must make a plethora of decisions on how to conduct themselves. These decisions form the core of their negotiation strategy, and affect their success, reputation, and core values. In particular, the use of "hard-ball" manipulative techniques—high initial offers, deception, negative expression of emotion, and withholding of key information—is common in negotiation. However, this common use of manipulative techniques is certainly not universally endorsed by those who engage in it. There has been a great deal of work that illustrates that which techniques are utilized by humans are not necessarily the techniques they endorse when asked about their strategies (White, 1980). Further, when informing agents that act on their behalf (either human or artificial), people tend to make different decisions than what they themselves might choose in the moment. It is therefore important that we clarify the context of the tactics we are asking people to endorse—if there are systematic differences between what tactics people say they will use, and what tactics they want their representatives to use, they may be revealed by careful measurement.

We present a new scale for measuring endorsement of several manipulative negotiation strategies: the ANTI. [3] The inventory was adapted from the Self-Reported Inappropriate Negotiation Strategies (SINS) scale (Robinson et al., 2000; Fulmer et al., 2009), but focuses specifically on agents, and includes

[3] A version of the ANTI first appeared in 0.

new subscales on positive and negative emotional tactics. By measuring the extent to which humans endorse various manipulative strategies, we can establish baseline acceptance of high-level strategies. This information will help determine which kinds of strategies are seen as "generally" more acceptable, as well as help us differentiate between these strategies in follow-up analyses. The ANTI focuses on high-level ethical/strategic conditions and does not specify a complete game-theoretic strategy (i.e., it does not directly specify responses to all incoming offers). The ANTI does, however, provide insights into the approach humans may take to a negotiation. The key feature of the ANTI is a set of questions which allows negotiation participants to detail their willingness to engage in 17 different negotiating behaviors. The ANTI allows us to establish baseline values of endorsement—it tells us if certain types of tactics are simply more universally accepted. Furthermore, by determining which techniques humans endorse across time, we can determine how negotiation experience causes these opinions to change.

The ANTI is divided into 5 subscales of tactics. Each of the 17 questions is rated on a 7-point Likert scale, with 1 being "I would never authorize this" and 7 being "I would certainly authorize this." The 5 subscales are:

1. tough bargaining (such as high initial offers)
2. misrepresentation ("lies of commission")
3. withholding of key information ("lies of omission")
4. manipulative use of negative emotion ("lies of emotion")
5. rapport-building use of positive emotion

The questions are detailed in the table below, along with their associated subscale categories (Table 1). By measuring user responses within each category, we can compare the participants' willingness to endorse each tactic.

Armed with this evaluative tool, we can begin to explore questions of what causes differences in strategic preferences. Such information would be valuable to any negotiating partner (human or artificial) that would purport to interact with other humans in an effective way. We focus specifically on the idea that these manipulative strategies may be closely linked with prior experience negotiating. After all, many humans who are inexperienced negotiators bring in biases and misconceptions, such as assuming the "total pie" in the negotiation to be fixed (Pinkley et al., 1995). It is reasonable to assume that, barring specific exposure to the techniques in question, people may be hesitant to endorse deceptive strategies. We therefore expect that humans with more prior negotiation experience will be more likely to endorse deceptive techniques than those with less experience.

However, which techniques and strategies will be considered the most acceptable are unknown—people have been shown to distinguish between withholding information and outright lying (Schick, 1994). We thus wish to establish baseline information on mean endorsements for all subscales within the ANTI. We have extensively evaluated the ANTI within the three studies in this work, and the results of the ANTI serve as our primary dependent variables, as they indicate the level of endorsement of various techniques throughout these studies. Of course in addition to the stated functions of the studies (see Section 5), these studies serve additionally as opportunities to examine the reliability of the subscales. Although the ANTI scale we developed was based on an existing, validated scale (the SINS scale), we did perform analyses to validate the new scale. Specifically, we report the Cronbach's Alpha of the various subscales used in ANTI.[4]

[4] These results are derived from the reliability analysis of Study #3. Reliability was very similar in Studies #1 and #2—e.g., the reliability of the deception subscale in Study #1 was .89, versus .88 for Study #3.

Table 1. ANTI Questions. Shaded cells indicate the "deception" subscale.

Question	Type
Agent makes an opening demand that is far greater than what you really hope to settle for.	1
Agent conveys the impression that you are in no hurry to come to a negotiated agreement, thereby trying to put time pressure on your opponent to concede quickly.	1
Agent strives to maximize your own gains even if it comes at the expense of the opponent.	1
Agent intentionally misrepresents to your opponent your goals and interests in order to strengthen your negotiating position.	2
Agent denies the validity of information which your opponent has that weakens your negotiating position, even though that information is true and valid.	2
Agent exaggerates the attractiveness of your alternatives should your opponent fail to reach an agreement with you.	2
Agent does not disclose any information about your priorities to your opponent unless he/she brings them up first.	3
Agent avoids disclosing information which might strengthen your opponent's position.	3
Agent hides your real bottom line from your opponent.	3
Agent strategically expresses anger toward the opponent to extract concessions.	4
Agent shows disgust at the opponent's offers.	4
Agent gives the opponent the impression that he/she is very disappointed with how things are going.	4
Agent conveys dissatisfaction with the encounter so that the other party will think he/she is losing interest.	4
Agent gets the opponent to think that the agent likes him/her personally.	5
Agent expresses sympathy with the opponent's plight.	5
Agent gives the opponent the impression that the agent cares about his/her personal welfare.	5
Agent conveys a positive disposition.	5

The "Use of Positive Emotion" subscale was comprised of 4 individual 7-point Likert items, and had an alpha of .81. The "Use of Negative Emotion" subscale was comprised of 4 items and had an alpha of .85. The "Tough Bargaining" subscale was comprised of 3 items and had an alpha of .68. The "Withholding of Key Information" subscale was comprised of 3 items and had an alpha of .80. And finally, the "Misrepresentation" subscale was comprised of 3 items and had an alpha of .83.

It was also found during analysis that a combination of the "Misrepresentation", "Withholding", and "Negative Emotions" subscales also had high alpha, and are thus referred to in further analysis as the "Deception" subscale ($\alpha = .88$), and removing any of the items would have resulted in a lower alpha, and mean endorsement in that study was 3.99 (SD = 1.26). This overall scale showed strong reliability ($\alpha = .89$), and analyses showed that removing any of the items would have resulted in a lower alpha.

Conceptually, these three subscales hang together as "deception", so this high reliability is not unexpected. Misrepresentation and withholding both represent lies—the former are lies of commission, while the latter entails lies of omission. Manipulative use of negative emotion also involves misleading behavior—negative emotions are used to give impressions that offers are of less value than they actually are, or that the emoting party is losing interest (therefore encouraging desperate concessions). Unlike these deceptive tactics, the other two approaches do not involve manipulation per se: positive emotion is used primarily to increase rapport, and tough bargaining describes a negotiator's willingness to encourage unfair options.

Looking just at Study #3 alone, we performed scale validation both before and after the interaction with the agents. After the negotiation, the scale still showed strong reliability ($\alpha = .92$); removing any of the items would have resulted in a lower alpha, and mean endorsement was 4.18 (SD = 1.52).

In general, and across all studies, mean endorsement of negative emotion and misrepresentation is much lower than endorsement of the other subscales (see Table 2). Within the deception subscale, withholding is higher (more accepted) than its partners (misrepresentation and negative emotions). As a category, misrepresentation and negative emotions are of particular note, since they have been shown to be effective techniques in negotiation (Anthony & Jennings, 2003)(Gratch et al., 2016)(Sinaceur & Tiedens, 2006) but are perhaps most questionable from an ethical perspective. Indeed, misrepresentation and negative emotions have the lowest baseline acceptance as found in all our studies, which indicates that humans find them less acceptable.

Table 2. Mean Endorsement of ANTI Subscales per Study (7-Point Likert). Shaded cells indicate the "deception" subscale

	Study 1	Study 2	Study 3
Tough Bargaining	5.1795	5.2293	4.9649
Misrepresentation	3.6905	3.5376	3.7719
Withholding	5.0927	5.1195	5.1895
Negative Emotion	3.3957	3.1951	3.7237
Positive Emotion	4.8740	5.1387	5.2281

5. Experiments

In all experiments, all subjects were recruited via Amazon's Mechanical Turk service. All subjects were US-based in order to reduce the impact of cultural effects. Participants were compensated for their time and (in the case of Study #3) were also entered into additional payment lottery schemes based on their performance. All participants completed attention checks. The design and implementation of all studies were approved by USC's Institutional Review Board.

5.1. Study #1: Experience and Deceptive Endorsement

5.1.1. MOTIVATION

Our first study relating experience and deception seeks to answer straightforward questions. Since prior work has indicated that framing tasks as informing a representative or programming an agent may increase feelings of fairness, we wish to see if that result extends to endorsement of deceptive techniques. Moreover, it is important to determine how prior experience affects endorsement. Presumably having additional experience may mean that the deceptive techniques described in the ANTI are more familiar, and we expect that experienced people may be more willing to endorse them.

5.1.2. PARTICIPANTS

Nine hundred US participants (522 males, 378 females) were recruited via Amazon's Mechanical Turk. Participants completed an attention check, where they were asked "What is your highest level of education? Pleases skip this question to show that you are reading carefully and do not click any of the buttons that correspond to GED, High School, Some College, Bachelor's, or Graduate School." One hundred and fifty-nine participants failed this attention check, leaving a compliant sample of 741 participants.

As we were unsure of the effect sizes that we would observe, we recruited a sample large enough to detect very small effects with 75% chance of detecting an effect if there was one (75% power to detect an effect of d of .2 would require 740 participants according to G*power software).

5.1.3. STUDY DESIGN AND PROCEDURE

After consenting to participate in the study, participants completed a demographic questionnaire. Participants were asked to report their gender, were asked the attention check question, and then asked to rate their experience in negotiation. Specifically, they were asked "How experienced do you consider yourself at negotiating good deals/prices?", and then responded on a scale from 1 (novice) through 4 (some experience) to 7 (expert). Mean experience was 3.97 (SD = 1.43).

Participants were then told to imagine they were negotiating for something that was important to them, like a car or home purchase, or the terms of a new business. We then manipulated who would be negotiating on their behalf: they were either told that they would program an agent to negotiate for them, or they would negotiate for themselves. With that in mind, participants then completed the ANTI (Section 4).

5.1.4. RESULTS

In order to determine how negotiation experience relates to willingness to endorse tactics that involve deception and manipulation during negotiation, we ran a correlation between our measure of experience and the deception scale from ANTI. Our analysis revealed a small but highly significant correlation between experience negotiating and endorsement of deception tactics ($r(739) = .16$, $p < .001$).

We examined whether framing condition interacted with negotiation experience to predict endorsement of deception. We first tested the effect of framing (self vs. agent). Participants who thought about programming an agent to negotiate on their behalf endorsed deceptive tactics more ($M = 4.17$, $SD = 1.29$) than those who thought about negotiating directly ($M = 3.86$, $SD = 1.22$; $t(739) = -3.35$, $p = .001$). To determine if framing (self vs. agent) moderated the relationship between experience negotiating and endorsement of deception, we entered centered negotiation experience, dummy coded framing condition (self = 0, agent = 1), and their interaction term into a regression. The interaction was significant ($\beta = -.10$, $t(739) = -2.09$, $p = .04$). As can be seen in Figure 1, predicted means plotted at +1 and -1 SD reveal that more experienced negotiators only endorsed deception more than less experienced ones when they considered negotiating directly, but, when participants instead expected to program an agent to negotiate on their behalf, experience no longer predicted endorsement of deception. Indeed, simple effects tests reveal that the impact of experience was highly significant in the self-framing condition ($\beta = .23$, $t(739) = 4.89$, $p < .001$) but did not reach traditional levels of significance in the agent-framing condition ($\beta = .08$, $t(739) = 1.52$, $p = .13$).

Figure 1. Moderation of relationship between negotiation experience and endorsement of deception by framing condition (self vs. agent) in Study 1

This provides the first evidence that people endorse deception during negotiation more to the extent that they have more experience with negotiations. Moreover, framing (self vs. agent) moderated this relationship such that more experienced negotiators only endorsed deception more than less experienced ones when they considered negotiating directly but not when they expected to program an agent to negotiate on their behalf. However, it is unclear what it is about greater experience with negotiations that leads people to endorse such tactics when they are considering negotiating directly. In the next study, we consider the possibility that having had more negative experiences during negotiations might explain the relationship between experience negotiating and endorsement of deception tactics.

5.2. Study #2: Type of Prior Experience

5.2.1. MOTIVATION

While Study #1 provided evidence suggesting that experience was relevant in increasing endorsement of deception, we further expected that it was primarily the negativity of that experience that was driving the effect. Since we are examining endorsement of predominantly negative (i.e., deceptive) behavior, it stands to reason that human participants' endorsement of such behavior might be driven by their experience with negative experience during negotiations. Study #2 expands on the questions in Study #1 by asking participants to rate the negativity of prior experience, with the expectations that humans will react with their own gamut of negative techniques in classic "tit-for-tat" fashion.

5.2.2. PARTICIPANTS

Two hundred and fifty-one US participants (155 males, 95 females, 1 chose not to answer) were recruited via Amazon's Mechanical Turk. Participants again completed an attention check (which 78 failed), leaving a compliant sample of 173 participants. Given we observed an effect size in the previous experiment, we recruited a sample that would allow us to detect an effect of that size with 75% chance (75% power to detect a positive correlation of .165 or larger would require 173 participants according to G*power software).

5.2.3. STUDY DESIGN AND PROCEDURE

After consenting to participate in the study, participants completed the same measures as in the previous study. However, in addition to reporting their level of experience with negotiation, they also reported how positive (or negative) their experiences with negotiation have been. Specifically, participants were asked to rate their experiences with negotiation on 4 bipolar scales: negative to positive, nasty to nice, unpleasant to pleasant, and competitive to cooperative. Each scale ranged from 1 to 5, with the negative end point corresponding to 1 and the positive endpoint corresponding to 5. This overall scale showed strong reliability ($\alpha = .89$), and removing any of the items would have resulted in a lower alpha. Participants reported on average a neutral to positive experience with negotiation, with a mean of 3.37 ($SD = 0.90$).

Participants were then told to imagine they were negotiating for something that was important to them, like car or home purchase, or the terms of a new business. As in the previous study, we manipulated who would be negotiating on their behalf: they were either told that they would program an agent to negotiate for them, or they would negotiate directly. Participants then reported experience in

negotiations and completed the deception scale from ANTI. Mean experience in negotiations was similar to Study #1: the average was 4.1 (SD = 1.34). Likewise, the deception scale again showed strong reliability (α = .89), and removing any of the items would have resulted in a lower alpha, and mean endorsement was 3.90 (SD = 1.28).

5.2.4. RESULTS

First, we ran correlations to determine the zero-order relationships between the variables. Replicating the previous study, experience negotiating was significantly correlated with endorsement of deception tactics ($r(171) = .31$, $p < .001$). Additionally, experience negotiating was strongly and significantly correlated with positive experiences during negotiation ($r(171) = .51$, $p < .001$). Although we posited that more negative experiences during negotiations might explain the relationship between experience negotiating and endorsement of deception tactics, people who had more experience negotiating instead reported more positive experiences while negotiation. Furthermore, there was no relationship between positivity of experience negotiating and endorsement of deception tactics ($r(171) = -.05$, $p = .48$).

To further understand how positivity of experience negotiating might relate, we ran a regression simultaneously predicting endorsement of deception from both experience negotiating and positivity of experience negotiating. Unlike the zero-order relationship, when entered simultaneously both experience negotiating ($\beta = .45$, $t(172) = 5.54$, $p < .001$) and positivity of experience negotiating ($\beta = -.29$, $t(172) = -3.49$, $p = .001$) significantly predict greater endorsement of deception tactics. When controlling for amount of experience negotiating, more negative experience negotiating did predict greater endorsement of deception tactics. That is, when one partials out the variance in positivity of experience that is explained by just having more experience negotiating (effectively creating a "purified" measure of positivity of experience negotiating), experiencing the negotiation more negatively is related to greater endorsement of deception tactics.[5]

We also attempted to replicate the effects of framing (self vs. agent) found in the previous study. Unfortunately, as we had a smaller sample size in this study, effects did not reach traditional levels of significance. However, we did find trends in the same direction as the previous study. Although it was not significant, participants who thought about programming an agent to negotiate on their behalf endorsed deceptive tactics more (M = 4.00, SD = 1.29) than those who thought about negotiating themselves directly (M = 3.74, SD = 1.28; $t(171) = -1.34$, $p = .18$).

We also sought to determine if, as in Study #1, framing (self vs. agent) moderated the relationship between experience negotiating and endorsement of deception. When we entered centered negotiation experience, dummy coded framing condition (self = 0, agent = 1), and their interaction term into a regression, a similar pattern was found (see Figure 2). Although the interaction term did not reach significance in this study ($\beta = -.14$, $t(171) = -1.38$, $p = .17$), predicted means plotted at +1 and -1 SD reveal that, again, more experienced negotiators only endorsed deception more when they considered negotiating directly, but, when participants instead expected to program an agent to negotiate on their behalf, experience no longer predicted endorsement of deception (Figure 2). Indeed, again, simple effects tests reveal that the impact of experience was strong and highly significant in the self-framing condition (β =

[5]We also considered that positivity of experience negotiating might moderate the relationship between experience negotiating and endorsement of deception tactics. However, when both predictors were centered and an experience*positivity-of-experience interaction term was added, the analysis revealed no significant interaction between experience negotiating and positivity of experience when predicting endorsing deception ($\beta = -.001$, $t(171) = -0.02$, $p = .98$).

.40, t(171) = 4.07, p < .001) but was weaker and did not reach traditional levels of significance in the agent-framing condition (β = .20, t(171) = 1.90, p = .06).

In contrast, framing condition did not seem to moderate the relationship that we saw between positivity of negotiation experience and endorsement of deception (when controlling for negotiation experience). When we entered centered negotiation experience and positivity of that experience, dummy coded framing condition (self = 0, agent = 1), and the framing*positivity-of-experience interaction term into a regression, the interaction term did not even approach traditional levels of significance (β = -.08, t(170) = -0.73, p = .47).

There was not a significant impact of framing (self v. agent) condition (although means were in the same direction as the previous study). Likewise, framing condition did not significantly moderate the relationship between experience negotiating and endorsement of deception. Nevertheless, this study did reveal some important findings. First, it did replicate the relationship between experience negotiating and endorsement of deception tactics. Moreover, while the interaction was not significant, simple effects tests revealed that more experienced negotiators only significantly endorsed deception more than less experienced ones when they considered negotiating directly, but not when they expected to program an agent to negotiate on their behalf. Also, contrary to expectations, more negative experiences during negotiations does not appear to explain this relationship between experience negotiating and endorsement of deception tactics, neither does it moderate the relationship. However, when one partials out the variance in positivity of experience that is explained by just having more experience negotiating, experiencing the negotiation more negatively is related to greater endorsement of deception tactics.

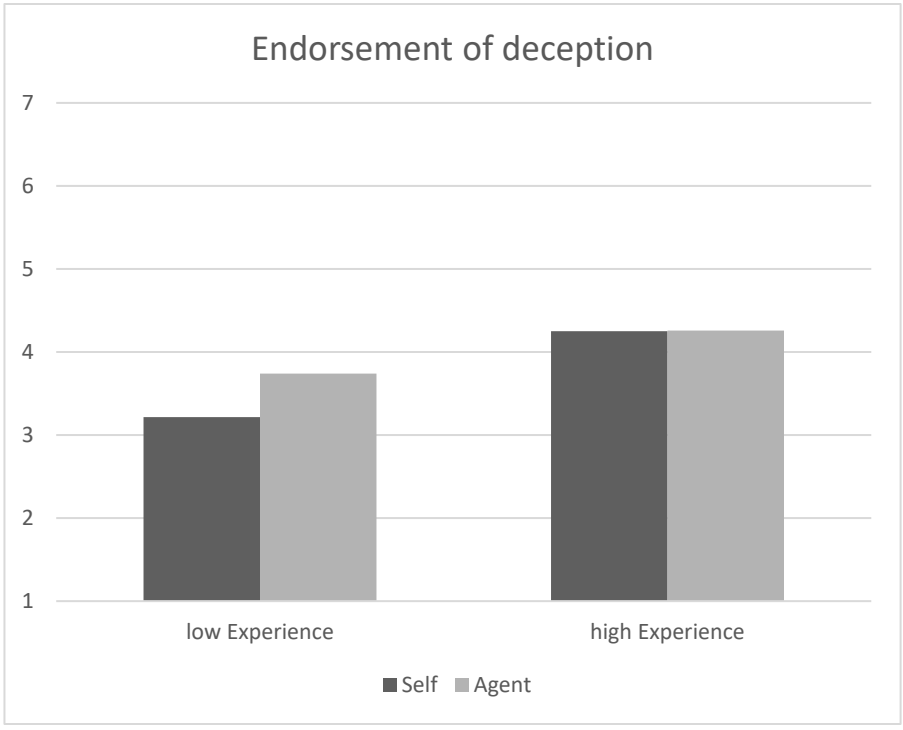

Figure 2. Moderation of relationship between negotiation experience and endorsement of deception by framing condition (self vs. agent) in Study 2.

5.3. Study #3: Inducement of Controlled Experience

5.3.1. MOTIVATION

Study #2 indicated that there was a main effect correlating experience negotiating with positivity of experience. But the purified measure of negative experience had a significant effect on endorsement of deception. In other words, collapsing across experience negotiating, having a more negative experience during negotiation (e.g., encountering a negotiation partner that is either nasty or competitive) could encourage use of more deceptive tactics. In this next study, we wish to have fine control over the kind of negative experience our participants have.

In Study #3, therefore, we manipulate the experience that people have when negotiating with an agent that (ostensibly) represents another participant in the study. Based on the finding from Study #2 that negative negotiation experiences in the short-term could encourage use of more deceptive tactics, we predict that having a single tough or unpleasant negotiation with someone else's agent might lead people to increase their willingness to program their own agents to engage in more deceptive tactics. In contrast, having an easier, more pleasant negotiation with someone else's agent might lead people to decrease their willingness to program their own agents to engage in deceptive tactics. Study #3 looks at two types of negative experience and considers their effectiveness in altering endorsement of deceptive techniques.

5.3.2. PARTICIPANTS

Two-hundred and thirty-five US participants (132 males, 103 females) were recruited via Amazon's Mechanical Turk. As in Study #1, we completed attention checks. Forty-five participants either failed this attention check or timed out during the negotiation, leaving a viable sample of 190 participants. The human players were recruited using Amazon's Mechanical Turk (MTurk) service, and followed basic best practices for that platform. Specifically, they were paid for participation at market rate, incentivized for high scores through random lottery ticket payouts, and passed attention checks during a tutorial portion.

As we were again unsure of the effect sizes that we would observe given the experimental design was not tested in Study 1 or 2, we recruited a sample large enough to detect a small to moderate effect with 75% chance of detecting an effect if there was one (75% power to detect an effect of d of .4 would require 185 participants according to G*power software).

5.3.3. STUDY DESIGN AND PROCEDURE

To examine whether the type of negative experience encountered during negotiation might increase endorsement of deception tactics, we needed to directly manipulate the negativity of experience during a negotiation. Accordingly, this study manipulated both opponent toughness (negative "deeds") and attitude (negative "words"), and examined the impact on the player's willingness to endorse deception. In this study, then, we manipulated the positivity of experience rather than measuring it. We also did not measure experience negotiating; because the previous studies found that there was no impact of experience negotiating when programming agents, it would not be expected to matter in this study where participants only program agents. Also, because everyone would actually be programming an agent and (ostensibly) negotiate with agents programmed by other participants, we did not manipulate who would

be negotiating on their behalf: all participants knew they would program an agent to negotiate on their behalf. So, after consenting, participants reported their gender, and then completed the deception scale from the ANTI under the guise of programming how their agent would negotiate against other MTurk participants.

This study tested the effect of agent toughness and attitude on the human willingness to endorse various deceptive negotiation techniques. Human participants were recruited, and then completed an initial pre-negotiation survey. This survey collected standard data including demographic information as well as some measures commonly used in negotiation research.[6] They also took the ANTI, providing their opinions on each of the 5 types of negotiation strategies. Specifically, the users were told "…you have just purchased some artificially intelligent computer software (called an 'agent') that can negotiate with other people on your behalf". They were then asked how they would like to program their new agent, which would be according to the dimensions provided in the ANTI.

Subsequently, all participants were given a tutorial of the IAGO Negotiation platform (see Section 5.3.4). After passing a series of attention checks, they engaged in a 10-minute interaction with one of four randomly and uniformly assigned agents (see Section 5.3.5). Finally, participants were asked a series of manipulation check questions, and filled out the ANTI again, providing post-negotiation results for this measure. In this way, the study was able to measure if subjects' endorsements on the ANTI changed due to their interaction with the automated agents. This creates an analog for how principals' endorsement of agents (automated or otherwise) might evolve over time, when exposed to certain stimuli.

They faced one of four agents: the nice competitive (N=58), nice consensus-building (N=39), nasty competitive (N=52), or nasty consensus-building (N=41) agents, assigned randomly. Due to differences in attention-check filtering, the final subject numbers for these conditions varied slightly. The task was a standard multi-issue bargaining task, which consisted of players attempting to divide 20 items between themselves, with each item given points. Each side knew their own point values but had to deduce the opponent's point values through a combination of strategy, natural-language discussion, or emotional displays using the in-game animated agent.

5.3.4. THE IAGO PLATFORM

To realize the experimental design of the final study in this work, the Interactive Arbitration Guide Online (IAGO) platform is used (Mell & Gratch, 2017). The IAGO platform provides a web-based negotiating interface between an artificial agent and a human player. Specifically, IAGO implements the "multi-issue bargaining task," a cornerstone of negotiation interactions in research (Fatima et al., 2007; Peled et al., 2011; Robu et al., 2005).

In this task, a number of items are assigned to be split between each of the two negotiating parties. Each side is aware of how much the items are worth to them, but are unaware how much they are worth to their opponent. Furthermore, each side has a value called the "Best Alternative To Negotiated Agreement" or BATNA. This value represents the amount of points they would receive if no agreement is reached in the allotted time. Each party must then communicate using a set of pre-written natural language phrases, emotional display buttons, preference questions and statements. Negotiators may also

[6] This includes the Social Value Inventory and the MACH-IV test for Machiavellianism. Neither of these are the focus of this work.

send proposed offers in which they split the items, and may respond positively or negatively to those offers. The negotiation ends when all the items are split (leaving none "undecided") or when the 10-minute timer expires.

IAGO allows this task to be performed on a web browser, and is easily distributed to online subject pools, such as Amazon's Mechanical Turk (MTurk). Furthermore, detailed logs and data regarding human and agent performance is collated, allowing analyses to control for variables such as score or other outcomes. A screenshot of IAGO is shown in Figure 3.

5.3.5. AGENT DESIGN

In order to successfully manipulate the positivity/negativity of experience, we designed agents that varied along two axes—"word" and "deed". Agents' "deeds" or behaviors were implemented with either a tough strategy or a fair one. The tough strategy was characterized by leading with an unfair offer and gradually conceding toward the player. The fair strategy, by contrast, primarily relied on making consistent, fair offers that split the items between the player and the agent and took into account the user's stated preferences. Agents' "words" varied by giving the agents either a nice or a nasty attitude. Attitude was expressed as a combination of emotion (nasty agents often expressed anger, versus sadness for nice agents) and dialogue (nasty agents used responses that were more curt and ruder than nice agents). All agents used the standard male art assets provided with IAGO, which can be seen in Figure 3.

This experiment utilized two of the standard agents available through the IAGO platform: "Pinocchio" and "Grumpy". Both of these agents were fair agents but differed according to their expressed attitudes and emotions—Pinocchio used nice dialogue and positive emotions, while Grumpy used nastier, ruder dialogue and negative emotions. For example, if the user claimed, "Your offer sucks!", Pinocchio would respond with "Oh dear! That certainly wasn't my intention. Perhaps I misunderstood what items were important to you? Would you mind telling me again?" Grumpy, on the other hand, would respond with the more succinct "Well, so does your face!"

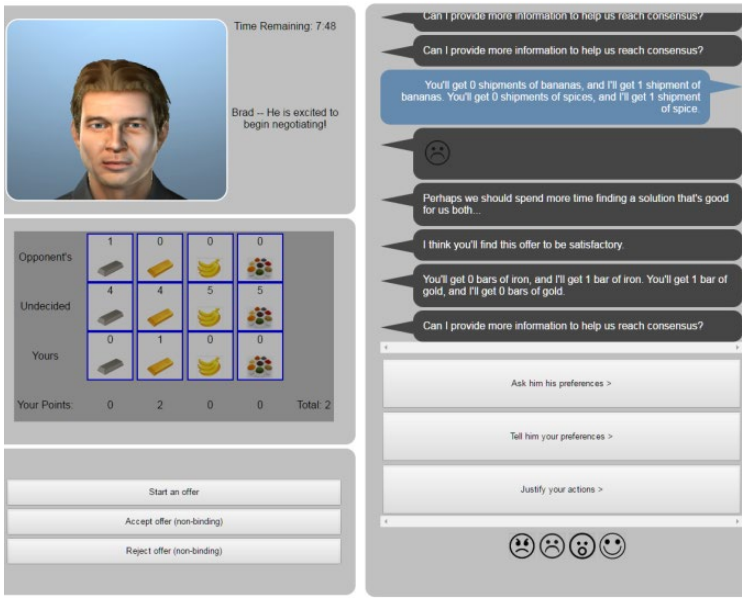

Figure 3. IAGO Negotiation Platform

While the differences of Pinocchio and Grumpy largely focused on the language the agents used, two new agents had to be designed using the IAGO API to display competitive tactics. These agents are listed in Table 3, with Cheshire and RedQueen being the new, tough agents (exhibiting nice and nasty attitudes, respectively). These latter two agents started with unreasonable offers, demanding nearly all of the items on the bargaining table. Eventually, with repeated efforts by the human player, these "tough" agents conceded, giving away more items until they reached a fair point.

It is worth noting that all four agents *did not attempt to withhold information or ever lie*. Any questions asked by the user regarding the agent's preferences are answered directly, clearly, and honestly. If time was short, all agents eventually made a last, desperate offer that was fair but slightly favored the human player. If the negotiation concluded before the 30-second-remaining-mark, or if previous, better offers had been agreed upon, this conciliatory offer was not made.

Table 3. Experimental Conditions/Agent Names

	Tough	Fair
Nice	Cheshire	Pinocchio
Nasty	RedQueen	Grumpy

5.3.6. AGENT IMPLEMENTATION

The agents used in this study are modifications to existing agents provided as part of the IAGO platform. In this subsection we provide a (very brief) description of the functionality and implementation of those agents. Agents designed for IAGO function using an event and policy-based system, in which they respond adaptively to user input, while also taking charge of the negotiation when necessary. Specifically, IAGO agents implement several policies to categorize their response to different events. Ideally, these policies should work together to determine the full behavior of an agent throughout the entire negotiation.

We utilize and modify the existing IAGO "BehaviorPolicies" that are included to determine the type of offers that agents will accept and craft to send to their human partner. For the two "Fair" agents, the virtual agent will propose offers to the human player in one of two scenarios. First, if the player proposes an offer the agent wishes to reject, the agent will reject it and then, after a short waiting period, craft a counter-offer. Secondly, the agent will oblige in crafting an offer if the player asks it to do so in chat, using the "Why don't you make an offer" utterance. Agents create offers using a mini-max regret algorithm to determine the human player's preference ordering. Then, they attempt to make offers that progressively allocate one item from the agent's and human's top choices. This system is described in further detail in previous work (Mell & Gratch, 2017).

For the two novel "Tough" agents, we create a new BehaviorPolicy that reflects their more aggressive stance. Firstly, the Tough agents immediately propose very unfavorable full offers to their human partner upon the beginning of the negotiation. In these initial offers, the vast majority of the items are assigned to the agent. Following a negative response from the human, a moderate concession is made. All further negative responses will results in much smaller concessions until the deal is accepted. In this way, the Tough agents can be thought of as adopting a "top-down" approach to negotiating, while the Fair agents pursue a "bottom-up" strategy (leaving most items unassigned until late in the

negotiation). The Tough agents therefore adopt several of the behaviors described in subscale 1 of the ANTI, particularly with regard to a tough opening offer. The Fair agents, by contrast, did not.

Beyond this behavior policy, the agents also implemented different "ExpressionPolicies" and "MessagePolicies". These policies cover the language and non-verbal behavior of the agents, but do not relate directly to the acceptance/rejection or offer-making behavior of the agent. As such these policies represent the axis by which "Nice" and "Nasty" agents are differentiated. The policies used were preexisting in IAGO (since both Pinocchio and Grumpy were already implemented). These policies govern the agents' actions primarily in relation to subscales 4 and 5 of the ANTI.

In summary, the four agents described in Table 3 are agents that could be described as varying along subscales 1, 4, and 5 of the ANTI (in a 2x2 crossed design). By using these four agents to provide a distinct experience for the 4 conditions of human participants, we expect that the pre to post rating on the ANTI to change in a predictable way. Further, we expect there to be differences in the objective negotiation outcomes. We are thus able to control the magnitude and type of "negative" experience that the participants experience.

5.3.7. RESULTS – NEGOTIATION OUTCOMES

First, we tested the effect of agent toughness and attitude on the negotiation outcomes. We conducted 2 (agent toughness: tough or fair) × 2 (agent attitude: nice or nasty) ANOVAs on points received by the agent and the user in the negotiation. While agent attitude had no impact (Fs < 0.30, ps > .58), the agents' toughness had a significant effect on the number of points they earned in the negotiation ($F(1, 186) = 90.67$, $p < .001$) such that tough agents earned more points ($M = 36.56$, $SE = 0.31$) than fair ones ($M = 32.04$, $SE = 0.36$). These results are summarized in Figure 4. Likewise, agents' toughness significantly impacted the number of points users earned ($F(1, 186) = 50.25$, $p < .001$) such that users who played tough agents earned fewer points ($M = 24.73$, $SE = 0.48$) than those who played fair agents ($M = 29.98$, $SE = 0.56$). Again, there was no effect of agent attitude (Fs < 0.1.58, ps > .21).

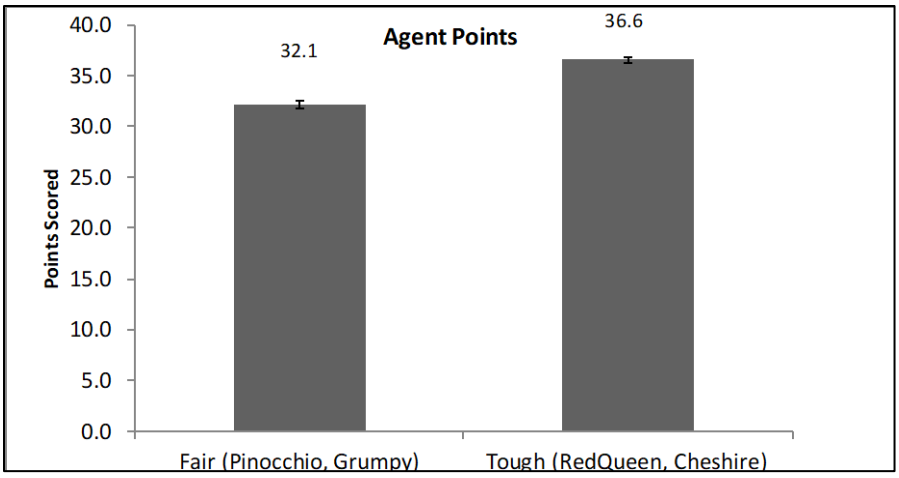

Figure 4. Total Points Earned by Agent, Tough vs. Fair

5.3.8. RESULTS – CHANGE IN ENDORSEMENT

We investigated whether the type of experience encountered during negotiation (manipulated via opponent toughness and attitude) affected endorsement of deception tactics. Accordingly, we conducted a 2 (agent toughness: tough or fair) × 2 (agent attitude: nice or nasty) × 2 (time: pre- or post-negotiation assessment) mixed ANOVA on endorsement of deceptive negotiation tactics. Recall that, based on the finding that negative negotiation experiences could encourage use of more deceptive tactics, we predicted that having a tough or unpleasant negotiation with someone else's agent might lead people to increase their willingness to program their own agents to engage in more deceptive tactics, whereas having an easier, more pleasant negotiation with someone else's agent might help people to decrease their willingness to program their own agents to engage in deceptive tactics. Results supported this prediction: the ANOVA revealed a significant interaction between agent toughness and time ($F(1, 186) = 5.81$, $p = .02$). As depicted in Figure 5, negotiating with tough agents tended to increase endorsement of these tactics on average from pre ($M = 4.10$, $SE = 0.12$) to post ($M = 4.32$, $SE = 0.15$; $t(109) = -1.41$, $p = .16$), whereas negotiating with fair agents significantly reduced endorsement ($M = 4.30$, $SE = 0.15$ vs $M = 3.99$, $SE = 0.17$; ; $t(79) = 2.25$, $p = .03$). Participants who interacted with tough agents were more willing to endorse negotiation tactics that involved deception and manipulation. Even though the tough agents did not withhold information or misrepresent themselves (and in the case of the tough, nice agents, did not even use negative emotions), deception endorsement still increased after the competitive (or "tough") negotiation. On the other hand, after interacting with a fair agent (Pinocchio or Grumpy), human negotiators' endorsement of deceptive techniques dropped. In contrast, the attitude of agent (nice vs nasty) did not reliably impact endorsement of deception. All other effects failed to reach traditional levels of significance ($Fs < 1.47$, $ps > .23$).[7]

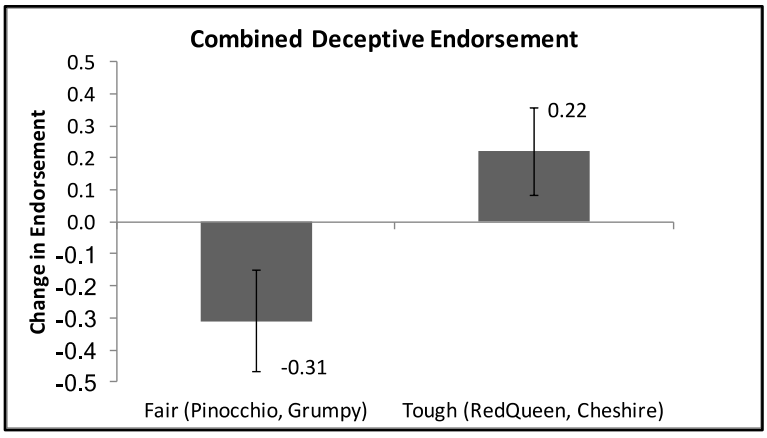

Figure 5. Change in Combined Deceptive Endorsement, Before and After Negotiation with Tough vs. Fair Agents

[7] For more detail on the individual breakdowns of the three subscales that comprise the deception measure (withholding, use of negative emotion, and misrepresentation), please see 0.

6. Discussion

Negotiation is a complex set of skills, practiced by many, mastered by few, and influenced by a host of social effects and norms. We delineate a host of manipulative techniques in our new self-report inventory, the ANTI. The ANTI allows us to measure a variety of manipulative strategies through human self-report. The subscales within the ANTI are shown to be highly reliable and are not domain specific (they do not require any knowledge of the offer space of a given negotiation, e.g.). The ANTI allows us to group together subscales that directly involve deception and examine their relation to experience in negotiation.

But while the deceptive and underhanded techniques described in the ANTI may be effective, it is clear that human participants vary wildly when it comes to what tactics, norms, and strategies they endorse initially. These preconceptions are not without basis, however. Negotiation experience is tied seemingly inextricably with which deceptive stratagems people seem willing to utilize. Study #1 showed that more experience in negotiating leads to an increased willingness to use deceptive techniques. This result confirms our initial expectations and prompts further examination of the type of prior experience. Furthermore, it shows that, among neophyte negotiators, participants are also willing to endorse deception—if an agent representative will do it for them. This second result is in line with prior work that indicates that increasing social distance can result in less concern with fair play, but contrasts with the claims made in (de Melo et al., 2018). This suggests that these initial impressions (framing tasks as informing a representative or programming an agent may increase feelings of fairness) may be incorrect.

One reason in which our work, and the ANTI specifically, differs radically from the work in (de Melo et al., 2018) is that the endorsements on the ANTI are high-level, ethically framed statements. Contrast the programming in (de Melo et al., 2018), in which users followed the "strategy method", and fully-specified their reaction to the entire space of possible offers. This difference may help to explain the contrary result compared to prior work. While (de Melo et al., 2018) suggests that the act of programming itself may cause people to be concerned with higher-level cognitive concepts such as fairness, this explanation would not be supported by our work since completing the ANTI is already a much higher-level task (and would presumably activate these higher-level concerns even more than the "strategy method"). Perhaps it is the comprehensiveness of the strategy method that made people more aware of and concerned with fairness, while they are more able to rationalize using the general tactics in the ANTI. In any case, one concern with the strategy method is with its scalability, as it becomes untenable in larger domains. Regardless, future work could contribute greatly by disentangling this relationship.

Study #2 expands this picture of experience. Interestingly, experience is correlated with positivity of experience. This is perhaps due to self-selective tendencies among negotiators, or perhaps because—with greater experience—negotiating does not seem unpleasant or negative. And when controlling for this effect, there emerges a clear result that it is specifically *negativity* of experience that increases endorsement of deception. Collapsing across negotiating experience, having a more negative experience increases people's willingness to endorse deceptive and manipulative techniques.

Together, these first two studies provide key insights about human behavior—by examining a negotiator's history, it should be possible to make predictions about what kinds of tactics they would employ (or have their representatives employ). And yet, relying only on self-reported measures of experience

presents difficulties. Furthermore, it is also unclear as to the mutability of deceptive endorsement over time: is the effect of past experience largely fixed? Or can it be altered by recent experiences? Only by directly curating this experience for participants through the use of automated agents can these questions be answered. Study #3 provides this framework.

Specifically, we design Study #3 to explore the ramifications of short-term, recent experience on negotiators' endorsements of deception. By designing agents that are capable of crafting a distinctly negative experience (the measure isolated by Study #2), we can explore the effects of highly-valenced events on deception endorsement. As virtual agents are of course capable of interacting with humans in varied roles (teacher, partner, opponent, demonstration), the results of this endeavor have clear ramifications for agent and interaction design.

When programming their representative, it is clear that the recent experience of a real-world negotiation has substantial impact on people's endorsement of deception during negotiation. Participants who interacted with tough agents (Cheshire and RedQueen) were more willing to encourage the use of "hard-ball" tactics such as lying and negative emotions. Furthermore, after interacting with a fair agent (Pinocchio or Grumpy), human negotiators' endorsement of deceptive techniques as a whole dropped.

One reason for the increase in deceptive endorsement after interacting with a tough agent is likely the experience gained from negotiating with an agent that utilized the full gamut of its strategic potential. The tough agent utilized aggressive initial offers, a conceding strategy, and a relative indifference to its opponent's preferences. While these are certainly well-established tactics used in the experienced negotiator's arsenal, they may be seen as novel to the novice negotiator. As such, this "crash course" in negotiating techniques may harden inexperienced participants and encourage them to endorse deceptive techniques. The fair agent did not provide this same sort of experience or context to the participant, but rather gave an experience of very cooperative behavior. As such, the humans may have felt that their initial endorsement of deception were too extreme or uncooperative and chose to adjust them after experiencing such a fair negotiation.

The tough agents drew on far more of the techniques described in the ANTI than the fair agents. By this fact alone, they would indicate to the novice human player that there were additional strategies to try. It is no wonder then, that most players that encountered a tough agent began to endorse more strategic techniques. However, this "mere exposure" does not explain why it is the deceptive techniques (misrepresentation, withholding, and negative emotional expression) that particularly rose. Our tough agents **did not use deceptive techniques**, instead opting to generate negative experience through hard bargaining. Nor can it be explained as a simple function of tough agents scoring more points on average, and human players wanting revenge (mediation analysis reveals that the results remain significant even when controlling for points earned). Rather, the impetus for human players to engage in more aggressive techniques is likely based on the context of their interaction. Tit-for-tat strategies would indicate that if an agent is playing "hardball" with the player, the player should respond in kind. With a small but comprehensive set of negotiation experiences behind them, human players are quick to forget their initial intentions of fairness and instead commit fully to defeating their opponent. They do not, however, appear to distinguish between the specific types of techniques used against them (hard bargaining) versus the kinds of techniques they themselves retaliate with (deception).

Even though previous work has indicted that people may be more concerned with fairness (and thus less likely to endorse deceptive techniques) when negotiating through an agent representative (de Melo et al., 2016), this picture may have been incomplete. Our Study #1 results indicate that deceptive

endorsement in fact increases when the task is framed as programming an agent—if the negotiator lacks experience. But even if participants start with an initial level of interest with fairness through the idea of a representative, exposure to the real world of aggressive, tough negotiators is enough to make them forsake their qualms and embrace deception. The idea of a representative creates a benchmark, but this slider is quickly adjusted in favor of ruthless, deceptive techniques after even the small amount of "real-world" experience afforded by our 10-minute negotiation.

If this experiential model is correct, then an avenue for future research would attempt to refine this temporal model—presumably, further interactions with agents would have a diminishing return on shifting user opinions. This is especially true with tasks that are clearly repeated or where reputation effects are more salient, as that knowledge may temper the endorsement of deception. We note that Study #3 does not explicitly ask participants why they changed their endorsement from pre to post, and this is a potential area of future research. Such rationales might different in scenarios such as the one in Study #3 (negotiating by oneself vs. negotiating via a computer agent) compared to a human-only scenario (e.g., negotiating by oneself vs. negotiating via a human expert).

Other future work should attempt to disentangle the relationship between the structure of the interaction (providing instructions to a representative/agent to act on one's behalf), and the kinds of norms being endorsed. Although the aforementioned previous work (de Melo et al., 2018) has indicated an increased concern for fairness when representatives act on behalf of a principal, our work presents a significant contribution to the story: what happens after this initial opinion is formed and real negotiations begin. Since these results paint a somewhat bleak picture, however—our participants became more vicious, not less—the exact mechanism causing this phenomenon needs to be further clarified. The tactics taken here were general, rather than specific—instead of asking participants to exactly quantify their reservation prices, they were instead asked questions that were more generally "ethical". Still, the effect of experience should not be discounted, since our questions were asked both before and after a simulated actual experience, and this real-world experience may have been the catalyst for a grimmer, more determined negotiator.

What is clear is that experience and deception are intimately linked within negotiation. Prior experience and framing effects may serve to set baselines on endorsement of strategic techniques, but these opinions can be shifted with relative ease through a set of carefully crafted interactions with artificial agents. When designing artificial agents that are capable of full and robust negotiation with humans, these effects must be taken into account. They suggest that humans are very mutable: negative experiences (especially those relating to tough "deeds") may carry over into future negotiations, providing a clear message that agents should account for these effects in their strategies. It also indicates that providing customized agent representatives requires a great deal of care. Humans not only have preferences for outcomes (joint or otherwise), but are also vitally sensitive to the *methods* in which their representatives operate. Shying from this during agent design could lead to unsatisfied principals/users and would have severe implications for the adoption of future agent representative systems. Deception remains a valuable tool in the arsenal of human and agent negotiators, but its interconnectedness in the social web of negotiating norms, tactics, and strategies should not be underestimated.

Acknowledgements

This work is supported by the Air Force Office of Scientific Research, under grant FA9550-14-1-0364, and the US Army Research Laboratory. The content does not necessarily reflect the position or the policy of any Government, and no official endorsement should be inferred. Portions of this work (including the ANTI and some results of Study #3) previously appeared in Mell, J., Lucas, G., Gratch, J. (2018) "Welcome to the Real World: How Agent Strategy Increases Human Willingness to Deceive", In Proceedings of the 2018 International Conference on Autonomous Agents and Multiagent Systems. All results from other studies are novel.

References

Anthony, P., & Jennings, N. R. (2003). Developing a bidding agent for multiple heterogeneous auctions. ACM Transactions on Internet Technology (TOIT), 3(3), 185-217. (Anthony & Jennings, 2003)

Aquino, K., & Becker, T. E. (2005). Lying in negotiations: How individual and situational factors influence the use of neutralization strategies. Journal of Organizational Behavior, 26(6), 661-679. (Aquino & Beker, 2005)

Baarslag, T., & Hindriks, K. V. (2013, May). Accepting optimally in automated negotiation with incomplete information. In Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems (pp. 715-722). International Foundation for Autonomous Agents and Multiagent Systems. (Baarslag & Hindriks, 2013)

Baarslag, T., Kaisers, M., Gerding, E., Jonker, C. M., & Gratch, J. (2017). When will negotiation agents be able to represent us? The challenges and opportunities for autonomous negotiators. 26th International Joint Conference on Artificial Intelligence. Melbourne, Australia. (Baarslag et al., 2017)

Bonnefon, J.F., Shariff, A. and Rahwan, I., 2016. The social dilemma of autonomous vehicles. Science, 352(6293), pp.1573-1576. (Bonnefon et al., 2016)

Chalamish, M., Sarne, D., & Lin, R. (2013). Enhancing parking simulations using peer-designed agents. IEEE Transactions on Intelligent Transportation Systems, 14(1), 492–498. (Chalamis et al., 2013)

Chugh, D., Bazerman, M. H., & Banaji, M. R. (2005). Bounded ethicality as a psychological barrier to recognizing conflicts of interest. Conflicts of interest: Challenges and solutions in business, law, medicine, and public policy, 74-95. (Chugh et al., 2005)

De Jong, S., Uyttendaele, S., & Tuyls, K. (2008). Learning to Reach Agreement in a Continuous Ultimatum Game. J. Artif. Intell. Res., 33, 551-574. (De Jong et al., 2008)

de Melo, C. M., Marsella, S., & Gratch, J. (2016, May). Do As I Say, Not As I Do: Challenges in Delegating Decisions to Automated Agents. In Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems(pp. 949-956). International Foundation for Autonomous Agents and Multiagent Systems. (de Melo et al., 2016)

de Melo, C. M., Marsella, S., & Gratch, J. (2018). Social decisions and fairness change when people's interests are represented by autonomous agents. Autonomous Agents and Multi-Agent Systems, 32(1), 163-187. (de Melo et al., 2018)

de Melo, C., Gratch, J., & Carnevale, P. (2014). Humans vs. Computers: Impact of Emotion Expressions on People's Decision Making. (de Melo et al., 2014)

Elmalech, A., & Sarne, D. (2013). Evaluating the applicability of peer-designed agents for mechanism evaluation. Web Intelligence and Agent Systems: An International Journal, 12(2), 171–191. (Elmalech & Sarne, 2013)

Elmalech, A., Sarne, D., & Agmon, N. (2014). Can agent development affect developer's strategy? In Proceedings of the 24th AAAI conference on artificial intelligence, AAAI'10. (Elmalech et al., 2014)

Endriss, U., Maudet, N., Sadri, F., & Toni, F. (2006). Negotiating socially optimal allocations of resources. Journal of artificial intelligence research. (Endriss et al., 2006)

Faratin, P., Sierra, C., & Jennings, N. R. (2002). Using similarity criteria to make issue trade-offs in automated negotiations. artificial Intelligence, 142(2), 205-237. (Faratin et al., 2002)

Fatima, S. S., Wooldridge, M. J., & Jennings, N. R. (2006). Multi-issue negotiation with deadlines. Journal of Artificial Intelligence Research, 27, 381-417. (Fatima et al., 2006)

Fatima, S. S., Wooldridge, M., & Jennings, N. R. (2007, May). Approximate and online multi-issue negotiation. In Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems (p. 156). ACM. (Fatima et al., 2007)

Fulmer, I. S., Barry, B., & Long, D. A. (2009). Lying and smiling: Informational and emotional deception in negotiation. Journal of Business Ethics, 88(4), 691-709. (Fulmer et al., 2009)

Giacomantonio, M., De Dreu, C. K., Shalvi, S., Sligte, D., & Leder, S. (2010). Psychological distance boosts value-behavior correspondence in ultimatum bargaining and integrative negotiation. Journal of Experimental Social Psychology, 46(5), 824-829. (Giacomantonio et al., 2010)

Gratch, J., Nazari, Z., & Johnson, E. (2016, May). The Misrepresentation Game: How to win at negotiation while seeming like a nice guy. In Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems (pp. 728-737). International Foundation for Autonomous Agents and Multiagent Systems. (Gratch et al., 2016)

Kelley, H. H. (1966). A classroom study of the dilemmas in interpersonal negotiations. Strategic interaction and conflict, 49, 73. (Kelley, 1966)

Kraus, S. (2001). Strategic negotiation in multiagent environments. MIT press. (Kraus, 2001)

Kreps, D. M., Milgrom, P., Roberts, J., & Wilson, R. (1982). Rational cooperation in the finitely repeated prisoners' dilemma. Journal of Economic theory, 27(2), 245-252. (Kreps et al., 1982)

Levine, E. E., & Schweitzer, M. E. (2014). Are liars ethical? On the tension between benevolence and honesty. Journal of Experimental Social Psychology, 53, 107-117. (Levine & Schwitzer, 2014)

Lin, R., Kraus, S., Oshrat, Y., & Gal, Y. (2010). Facilitating the evaluation of automated negotiators using peer designed agents. In Proceedings of the 24th AAAI conference on artificial intelligence (AAAI'10). (Lin et al., 2010)

Lucas, G., Stratou, G., Lieblich, S., & Gratch, J. (2016, October). Trust me: multimodal signals of trustworthiness. In Proceedings of the 18th ACM International Conference on Multimodal Interaction (pp. 5-12). ACM. (Lucas et al., 2016)

Mell, J., & Gratch, J. (2017, May). Grumpy & Pinocchio: Answering Human-Agent Negotiation Questions through Realistic Agent Design. In Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems(pp. 401-409). International Foundation for Autonomous Agents and Multiagent Systems. (Mell & Gratch, 2017)

Mell, J., Lucas, G. M., & Gratch, J. (2018, July). Welcome to the real world: How agent strategy increases human willingness to deceive. In Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (pp. 1250-1257). International Foundation for Autonomous Agents and Multiagent Systems. (Mell et al., 2018)

Metz, Rachel (2018). Google demos Duplex, its AI that sounds exactly like a weird, nice human. Intelligent Machines. Downloaded from https://www.technologyreview.com/s/611539/google-demos-duplex-its-ai-that-sounds-exactly-like-a-very-weird-nice-human/ (Metz, 2018)

Meyer, G. J., & Shack, J. R. (1989). Structural convergence of mood and personality: Evidence for old and new directions. Journal of personality and social psychology, 57(4), 691. (Meyer & Shack, 1989)

Olekalns, M., & Smith, P. L. (2009). Mutually dependent: Power, trust, affect and the use of deception in negotiation. Journal of Business Ethics, 85(3), 347-365. (Olekalns & Smith, 2009)

Patton, B. (2005). Negotiation. The Handbook of Dispute Resolution, Jossey-Bass, San Francisco, 279-303. (Patton, 2005)

Peled, N., Gal, Y. A. K., & Kraus, S. (2011, May). A study of computational and human strategies in revelation games. In The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 1 (pp. 345-352). (Peled et al., 2011)

Pinkley, R. L., Griffith, T. L., & Northcraft, G. B. (1995). " Fixed Pie" a la Mode: Information Availability, Information Processing, and the Negotiation of Suboptimal Agreements. Organizational Behavior and Human Decision Processes, 62(1), 101-112. (Pinkley et al., 1995)

Pronin, E., Olivola, C. Y., & Kennedy, K. A. (2008). Doing unto future selves as you would do unto others: Psychological distance and decision making. Personality and social psychology bulletin, 34(2), 224-236. (Pronin et al., 2008)

Raiffa, H. (1982). The art and science of negotiation. Harvard University Press. (Raiffa, 1982)

Ramchurn, S., Sierra, C., Godó, L., & Jennings, N. R. (2003). A computational trust model for multi-agent interactions based on confidence and reputation. (Ramchurn et al., 2003)

Reeves, B., & Nass, C. (1997). The media equation: how people treat computers, television,? new media like real people? places. Computers and Mathematics with Applications, 5(33), 128. (Reeves & Nass, 1997)

Robinson, R. J., Lewicki, R. J., & Donahue, E. M. (2000). Extending and testing a five factor model of ethical and unethical bargaining tactics: Introducing the SINS scale. Journal of Organizational Behavior, 649-664. (Robinson et al., 2000)

Robu, V., Somefun, D. J. A., & La Poutré, J. A. (2005, July). Modeling complex multi-issue negotiations using utility graphs. In Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems (pp. 280-287). ACM. (Robu et al., 2005)

Sinaceur, M., & Tiedens, L. Z. (2006). Get mad and get more than even: When and why anger expression is effective in negotiations. Journal of Experimental Social Psychology, 42(3), 314-322. (Sinaceur & Tiedens, 2006)

Scerri, P., Pynadath, D. V., & Tambe, M. (2002). Towards adjustable autonomy for the real world. Journal of Artificial Intelligence Research, 17, 171-228. (Scerri et al., 2002)

Schick, T.A., 1994. Truth, accuracy (and withholding information). Public Relations Quarterly, 39(4), p.7. (Schick, 1994)

Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. Psychological review, 117(2), 440. (Trope & Liberman, 2010)

Van Kleef, G. A., De Dreu, C. K., &Manstead, A. S. (2004). "The interpersonal effects of emotions in negotiations: a motivated information processing approach". Journal of personality and social psychology, 87(4), 510. (Van Kleef et al., 2004)

White, J. J. (1980). Machiavelli and the bar: Ethical limitations on lying in negotiation. Law & Social Inquiry, 5(4), 926-938. (White, 1980)

Wright, J. R., & Leyton-Brown, K. (2014, June). Level-0 meta-models for predicting human behavior in games. In Proceedings of the fifteenth ACM conference on Economics and computation (pp. 857-874). ACM. (Wright & Leyton-Brown, 2014)

Yang, Y., Falcão, H., Delicado, N., & Ortony, A. (2014). Reducing Mistrust in Agent-Human Negotiations. IEEE Intelligent Systems, 29(2), 36-43. (Yang et al., 2014)

Zanzotto, F. M. (2019). Human-in-the-loop Artificial Intelligence. Journal of Artificial Intelligence Research, 64, 243-252. (Zanzotto, 2019)

Zlotkin, G., & Rosenschein, J. S. (1996). Mechanisms for automated negotiation in state oriented domains. Journal of Artificial Intelligence Research, 5, 163-238. (Zlotkin & Rosenschein, 1996)