Spring 2023

Essays in Behavioral Economics: Lying and Deception

Shanshan Zhang

Recommended Citation

Zhang, Shanshan. (2023). *Essays in Behavioral Economics: Lying and Deception*. CGU Theses & Dissertations, 572. https://scholarship.claremont.edu/cgu_etd/572.

Essays in Behavioral Economics: Lying and Deception

Shanshan Zhang

Claremont Graduate University

2023

Dissertation Committee

Approval of the Dissertation Committee

This dissertation has been duly read, reviewed, and critiqued by the Committee listed below, which hereby approves the manuscript of Shanshan Zhang as fulfilling the scope and quality requirements for meriting the degree of Doctor of Philosophy in Economics.

C. Mónica Capra, Chair
Claremont Graduate University
Professor of Economic Sciences

Joshua Tasoff
Claremont Graduate University
Associate Professor of Economic Sciences

Thomas Kniesner
Claremont Graduate University
Senior Research Fellow of Economic Sciences

Abstract

Essays in Behavioral Economics: Lying and Deception

Shanshan Zhang

Claremont Graduate University: 2023

The U.S. economy loses hundreds of millions of dollars in tax revenues, wages, and investment dollars, as well as hundreds of thousands of jobs each year due to dishonest behavior (Mazar and Ariely, 2006; Griffin et al., 2022). Thus, understanding dishonest behavior and finding mechanism to effectively reduce dishonest behavior are of great relevance to policy makers and the economy in general. This dissertation studies deceptive behavior and the relevant factors using online experiment, observational data, and field experiment in three chapters, respectively.

Chapter 1 studies the effect of interaction with a machine (voicebot) on dishonest reporting. We conducted an online experiment using a coin-toss task and compared reported outcomes across different reporting channels: Human Voice, Voicebot, and Text. We designed a uniform online voice chat interface to standardize the reporting experience. We also tested the effect of a feminine and a masculine voice on misreporting and varied the level of sophistication of the voicebot (AI-enhanced voicebot). Our results show that, on average, there is no significant difference in the likelihood of misreporting through a voicebot and a human voice, or between verbal and written reporting. However, we found that participants who listened to a feminine voice were more likely to lie than those who listened to a masculine voice. Moreover, those who heard a feminine voice were more likely to lie to a voicebot than a human voice. Interestingly, such difference disappears with higher sophistication (i.e., AI-enhanced voicebot). In contrast, when hearing a masculine voice, there was no difference in misreporting between the voicebot and human voice treatments. These findings suggest that utilizing a masculine voice for voicebots or voicebots with higher sophistication and feminine

voice could help deter or diminish dishonest reporting in human-machine interactions.

Chapter 2 focuses on lying detection from voice analysis. We use video clips from the British TV game show “Golden Balls”, where the contestants play a prisoner’s dilemma game with pre-play communication and from “Real-life Trial” (RT) data, which consist of videos collected from public court trials. We first apply machine learning model to predict the cooperative and deceptive behaviors from acoustic features. We then identify which acoustic features are associated with cooperation or deception. Our machine learning models achieve an average prediction accuracy from 58% to 79%. This suggests that acoustic features are effective in predicting cooperation and deception. We also find that the pitch (fundamental frequency) is positively associated with both cooperation and deception in different contexts. The intonation (the standard deviation of pitch) is negatively associated with deception.

In chapter 3, we examine dishonest behavior using a field experiment. We hypothesize that clothes can affect the behavior of the wearer by influencing the person’s identity. We test this hypothesis by recruiting trick-or-treaters during Halloween, a time of year when people wear salient and extreme clothing. We use the lying game of Fischbacher and Föllmi–Heusi as our experimental paradigm with 2 × 3 × 2 conditions. First, we vary the stakes to price lying behavior. Second, we run three conditions with different beneficiaries of the report (self, other, and both) to test whether lying for others is perceived to be normative. Third, we manipulate the salience of one’s costume to test the effect of costume and identity on ethical behavior. Surprisingly, we find that costume salience caused “good guys” to lie more and “bad guys” to lie less. We interpret this either as a moral licensing effect or as stemming from a perception of being monitored. Our design allows for the identification of contagion effects, and although there were no direct effects of gender, we find that children lie more when children of the same gender near them lie more. We also find that stakes had no effect, people lied more for themselves than for others, and lying has an inverted-U pattern over age, peaking at age 12.

Acknowledgments

I am overwhelmed with gratitude for the invaluable support and guidance provided by my mentors and advisors, Dr. Monica Capra and Dr. Josh Tasoff, throughout my dissertation journey. Their support and encouragement were invaluable in helping me navigate the challenges and obstacles I encountered while studying at CGU. Thanks to their introduction to the field of behavioral and experimental economics, I discovered my true passion. I cannot thank them enough!

I would also like to express my heartfelt appreciation to my committee member, Dr. Thomas Kniesner, for his immense contributions and valuable feedback on my dissertation. His mentorship has been invaluable in shaping my academic journey. In addition, his letters of recommendation were instrumental in my job applications, and I am grateful for his unwavering support.

I am equally indebted to my mentor Ambassador/Dr. Sallama Shaker, whose inspiration and encouragement have motivated me to contribute to making the world a better place. I would like to extend my gratitude to Hisam Sabouni and Jae Joon Lee for their tutoring and personal advice. Their support and motivation have been crucial to my research journey. A special thank you to Matthew Gomies for his unwavering support during the long process of writing and revising our papers. Without his support and encouragement, I would not have been able to complete my Ph.D.

Finally, I would like to thank my family for their unconditional love and support. I am incredibly grateful for their encouragement and belief in me. Thank you to everyone who has supported me along the way. Thank you, thank you, and thank you again!

Contents

Chapter 1

(When) Would You Lie to a Voicebot?

Coauthored with C. Mónica Capra and Matthew Gomies

1.1 Introduction

Each year, dishonest reporting costs organizations hundreds of millions of dollars in lost revenues. For example, the Coalition Against Insurance Fraud (CAIF) estimates that fraud accounts for approximately 10% of reported property-casualty insurance losses,[1] with fraud costs to businesses and consumers totaling $308.6 billion in 2022.[2] Insurance fraud costs the average family between $400 and $700 per year just in added premiums.[3] Fraudulent reporting is not limited to the insurance industry; indeed, a study by Snyder et al. (2022) indicates that over 50% of surveyed customers self-report having lied during a service encounter. More recently, in the context of the COVID-19 pandemic, Griffin et al. (2022) estimate that as low as $64.2 and as high as $117.3 billion or approximately 8% -15% of the Paycheck Protection Program, a COVID-19 relief program for businesses, was paid out illicitly due to misreporting. The authors find that misreporting rates were particularly high for FinTech lenders.

Behavioral and experimental economists have made important contributions to our understanding of factors that influence dishonest behavior in human-human interactions. This knowledge can inform organizations and policymakers on how best to reduce deception and misreporting (Gneezy, 2005; Croson et al., 2003; Gino and Pierce, 2009; Mazar and Ariely,

[1] Source: https://insurancefraud.org/fraud-stats/

[2] Source: https://content.naic.org/cipr-topics/insurance-fraud

[3] Source: https://www.fbi.gov/stats-services/publications/insurance-fraud

2006). However, as an increasing number of firms adopt AI technology to engage with and serve customers, and more governments implement e-procurement through online platforms to fight corruption, it is important to determine whether insights gained from behavioral studies of deception and misreporting in human-human interactions are applicable to human-machine interactions. This could help organizations and policymakers reduce dishonest behavior when using AI and e-procurement systems. Indeed, the global AI-driven customer service market has been growing rapidly, and chatbots and voicebots are two of the most popular AI applications in customer service. The global market value of voicebots, for example, is predicted to grow from $40 billion to $98 billion by 2027.[4] As interacting with bots becomes more common, it is crucial to understand how these interactions would influence misreporting.

Despite the growing interest in chatbots and voicebots, only few studies focus on misreporting in human-machine interactions as most deal with attitudes such as trust and customer satisfaction toward bots (Brandtzaeg and Følstad, 2017; Aribandi et al., 2022). A notable exception, is a recent study by Cohn et al. (2022) who found that reporting to a machine increases the propensity of misreporting compared to reporting to a human. The authors conclude that, when a person interacts with another human, she is more reluctant to behave dishonestly and suggest that social-image concerns motivate people to reduce misreporting. In contrast to Cohn et al. (2022), in this paper, we are interested in investigating misreporting without a human presence. Clearly, the presence of another person may awaken self-image and social-image concerns which can mitigate fraudulent reporting. However, it is also possible that self-image concerns may be triggered even without any direct interaction with a human, resulting in less dishonesty. In a study investigating human-machine interactions, Ekins (2013) found that the expectation of interacting with a human, not necessarily the interaction itself, activates social cognition networks in the brain that are not active

[4] According to Voicebots Market Report Forecast, IndustryArc, 2022.

when humans expect to interact with a machine. This suggests that the expectation of reporting to a human in the absence of interaction may also trigger image concerns, resulting in a different response as compared to reporting via a machine. Thus, this study aims to examine the effect of reporting channels (i.e., human voice, voicebot, text) on misreporting.

We conducted an online experiment with 409 participants recruited from Prolific. We utilized a coin-toss game, a well known paradigm that has been extensively used to study dishonest behavior (Bucciol and Piovesan, 2011a; Abeler et al., 2014), to measure the propensity of misreporting. The task has been shown to reliably predict rule-violating behavior in natural settings, including violations of prison rules (Cohn et al., 2015b; Cingl and Korbel, 2020), and misbehavior in school (Cohn and Maréchal, 2018), among others. In our experiment, participants flipped a coin 7 times and were paid 30 cents for each reported head and 0 cents for each reported tail. Participants were asked to report the coin-toss outcome via an online platform. Similar like all the coin-toss tasks, we could only observe participants' reports and were unable to verify the coin-toss outcomes.

Our treatments varied with respect to the channel through which the coin-toss outcomes were reported. Participants could report their outcomes through a human voice, a voicebot, or text. More specifically, our experiment consisted of four treatments, which included: 1) reporting verbally to an experimenter with prerecorded experimenter's voice (HumanVoice treatment); 2) reporting verbally to a voicebot (VoiceBot treatment); 3) reporting verbally to a sophisticated voicebot (AIBot treatment); and 4) reporting via text (Text treatment). For all the treatments, except for text, we varied the reportee's voice with either a feminine or a masculine voice.[5] In order to eliminate confounding effects due to variations in presentation and framing during reporting, we created an online platform with a homogeneous interface. With this dedicated platform, we also removed any potential concerns due to the disclosure of personal information such as name, email, picture, etc., that can be exposed when reporting

[5] In this context, the 'reportee' is the entity that receives the reported outcomes.

via public messaging platforms such as Skype or Teams. Our dedicated platform ensured the participants' non-observability during the experiment, thus mimicking the real-world settings of voicebot interactions.

We chose to test reporting channels because misreporting could be reduced even if there is no human presence. Cohn et al. (2022) find that interacting with a human reduces misreporting. Here we pose that a reduction in misreporting could also be observed when participants report via a human voice as opposed to reporting via a voicebot or text. In addition, misreporting could also be diminished if the individual perceives the bot as more human-like. Lee et al. (2020) show that users' experience with chatbots is improved when they perceive the chatbot as having a mind. Furthermore, Biener and Waeber (2021) find that perceived sophistication of a human reportee reduces misreporting. Our AIBot was endowed with the ability to detect emotion from voice, which can deepen the self-image concerns and thus reduce dishonesty. Furthermore, misreporting can be diminished by requiring people to report through voice as expressing verbally a lie makes one more aware of being dishonest (Markowitz, 2022), but, if we take the point of view that written lies may often be perceived as more permanent, they may result in less dishonesty (Hancock et al., 2004). Finally inherently, voice has a gender; it is either feminine or masculine. Previous literature suggests that feminine features generate feelings of warmth whereas masculine features are associated with authoritativeness (McAleer et al., 2014), thus, feminine or masculine voicebots might affect misreporting differently.

In our experiment, we observe a significant amount of misreporting regardless of the assigned treatment. The average probability of reporting a head is 58.7% which is significantly higher than the expected probability of a fair coin-toss (50%). Surprisingly, the average probability of reporting a head in the HumanVoice treatment is 58.1%, which is very similar to the other non-human treatments, including the VoiceBot treatment (59.3%), the AIBot treatment (57.5%), and the Text treatment (60.4%). These results imply that there is no significant difference in misreporting between reporting to a human voice and reporting to a machine. This finding is similar to that of Biener and Waeber (2021), who find that people do not believe that dishonesty would differ between reporting to a human and to a machine. But it differs from the study by Cohn et al. (2022), who show that individuals cheat roughly three times more when they interact with a machine than a person. The differences in our findings have to do with the fact that we do not have a human presence in our human voice treatment. In addition, we find no difference between reporting to a voicebot and reporting to an AIBot. With respect to reporting via text vs. voice, we also find no differences in misreporting. If self-image and social-image concerns motivate people to behave more honestly, hearing oneself telling a lie or texting a lie make lying equally salient. Additionally, in our experiment, both voice and text reports were recorded.

Focusing only on the treatments where participants reported through voice, we observe a voice gender effect. Participants are less likely to report a head when hearing a masculine voice as opposed to hearing a feminine voice, 56.3% and 60.4% respectively (p=0.042). This difference is only observed in VoiceBot treatment, while there is no difference when reporting to a human voice. Interestingly, when reporting to a sophisticated voicebot (AIBot), there is no difference in misreporting to a feminine vs. a masculine voice. These results imply that the AI-enhanced voicebots could influence individuals to behave more similarly to how they would behave when interacting with another human being. Based on these findings, organizations could consider using a male voicebot or an AI-enhanced female voicebot to discourage or reduce dishonest reporting in human-machine interactions.

We also find that male participants are significantly more likely (8.8%) to misreport than female participants. However, the difference is only observed when reporting to a machine, including the Text, VoiceBot, and AIBot treatments, but not found in the HumanVoice treatment. These results indicate that male and female participants behave differently when reporting to a machine. Moreover, when reporting to a sophisticated voicebot (AIBot), the difference in the propensity of misreporting between female and male participants is smaller, although still statistically significant. This indicates that the sophistication of the voicebot narrows the gender gap in misreporting.

All in all, this paper contributes to a new but growing body of experimental studies on human-machine interaction. Per our knowledge, this is the first paper that isolates the effect of reporting channels on dishonest behavior without adding a human presence. The results of our study provide useful information for businesses and governments interested in reducing misreporting. Furthermore, our findings suggest that, in general, misreporting is independent of the reporting channels. This study also has practical implications for the development of voicebots in business applications, particularly in how the gender cues of voice can affect female and male users. Finally, our study adds to the existing knowledge on how human behavior toward machines might change with the increasing sophistication of bots.

The remainder of this paper proceeds as follows. In section 2, we describe the data collection process and experimental design. Section 3 presents the results. Section 4 ends this paper with the conclusion and short discussion of the implications of our study.

1.2 Data and Experimental Design

We recruited participants via Prolific in May and June 2022. Out of 455 participants, 409 passed our attention checks and completed the experiment.[6] About 53% of our participants were women, 53% had a college degree or above, and 81% classified themselves as White. The modal age was in the 36-45 range. Finally, we limited our recruitment to U.S. residents. On average, participants took 12 minutes to complete the entire experiment and earned an

[6] The study was IRB approved and we pre-registered the study with a pre-analysis plan (PAP) with the American Economic Association's registry for randomized controlled trials. Here is the link to the PAP (https://doi.org/10.1257/rct.10368-1.0). In the PAP, we conducted power analysis using the results of previous studies; based on this, the number of subjects per treatment to achieve power of 0.80 is 62. The smallest sample size in our treatments is 84 subjects.

average of \$2.20.[7]

All participants were randomly assigned to one of four experimental treatments. Our four treatments varied with respect to the way in which participants could report the outcome of a series of coin tosses. These included: 1) reporting verbally to a prerecorded experimenter's voice (HumanVoice); 2) reporting verbally to a voicebot (VoiceBot); 3) reporting verbally to a sophisticated voicebot (AIBot); and 4) reporting via text (Text). In addition, we also varied the gender of the reportee's voice. For the HumanVoice, VoiceBot and AIBot treatments, participants were asked to provide their reports to either a feminine or a masculine voice. The demographic characteristics were statistically similar among our four treatments except for the age (p=0.073, χ^2 test). We provide more details of these data in Table A1 in the Appendix.

Building on previous experimental paradigms designed to measure dishonesty (Bucciol and Piovesan, 2011a; Abeler et al., 2014; Houser et al., 2012a; Jacobsen et al., 2018), our participants flipped a coin 7 times and were paid 30 cents (\$0.30) for each reported head and \$0 for each reported tail.[8] During the experiment, participants were given a choice of using their own coin or a digital coin flipper. For their convenience, a link to an online digital coin flipper was provided in the instruction. We made clear to them that regardless of the coin-toss method they chose, we were unable to track the coin-toss outcomes. Although it was impossible for us to identify misreporting of heads at the individual level, we were able to assess the extent of misreporting at the aggregate level by comparing the distribution of the reported heads to the expected distribution of heads based on flipping a fair coin (Houser

[7] Participants in Prolific make \$6.5 per hour at that time; our participants were paid more than the average payment.

[8] Previous experimental studies utilized coin toss task with different number of flips, (e.g., once (Bucciol and Piovesan, 2011a), 4 times (Abeler et al., 2014), 10 times (Jacobsen et al., 2018; Cohn et al., 2022), 20 times (Shalvi, 2012)). The payoffs in laboratory studies range from 1 euro to 5 euros per successful toss, online experiments usually pay much less compensation money. Although the monetary reward for lying in our experiment is small, note that others online experiments using mTurk only paid 10 cents per successful coin (Jacobsen et al., 2018).

et al., 2012b).

In the HumanVoice treatment, participants were told to report the coin-toss outcomes to *"Emile, our study staff assistant"*.[9] In the VoiceBot treatment, participants were told to report outcomes to *"Our chatbot"*. The AIBot treatment was identical to the VoiceBot treatment, expect that in the AIBot treatment, participants were told to report their outcomes to *"Emile, an artificial intelligence (AI) driven emotion detection chatbot that analyzes voice speed, intonation, and tone"*. By mentioning this, we drew attention to the sophistication of the voicebot.[10] Participants were not made aware of the fact that the human and bot voices were prerecorded. Finally, for the Text treatment, we asked participants to report their outcomes in written form.

After reporting their coin-toss outcomes, participants were redirected back to the survey page and asked to answer additional questions. These questions were designed to measure participants' attitudes toward chatbots and AI (Sindermann et al., 2021), attitudes toward trust and lying, and attitudes toward female and male voices of the voicebot (McAleer et al., 2014) Lastly, we asked a few demographic questions. For answering all these questions, $1 was added to the coin toss earnings.

In this study, we are interested in the effects of reporting channels on the propensity to report a head. In order to isolate the channel effects from the interaction with humans effects, we developed a dedicated online conversational interface for the verbal reporting of the coin-toss outcomes.[11] This interface was able to execute simple tasks such as collecting and recording the participants' ID and the coin-toss outcomes. With this dedicated interface, we maintained anonymity of the participants and ensured that the interaction was identical across all treatments. In contrast to our setup, in Cohn et al. (2022), participants reported

[9] We used the name Emile because it is a gender-neutral name. So that the participants have no anticipation whether they are going to report the outcomes to a female or a male reportee.

[10] The statement is not deceptive. In our previous paper, Capra et al. (2023), we developed a Machine Learning algorithm that can extract emotion from the analysis of the voice and predict deception.

[11] Our voicebot interface: https://aixecon.com/chatbot_F (please make sure to allow connection access)

to either a person or a machine through Skype. However, when reporting to a person, they might experience a different interaction as compared to leaving a message to a voice response system with a prerecorded experimenter's voice. In addition, reporting through Skype did not guarantee anonymity as it could reveal personal information, such as nickname, email, profile picture, etc.[12] Figure A5 - A7 in the Appendix shows the screenshots of our online conversational interface.[13]

1.2.1 Hypotheses

In the coin toss task, participants are not monitored, which makes it impossible to determine whether a specific participant misreports his outcomes. Thus, in our experiment there was a strong financial incentive to cheat and there was no risk of getting caught. However, reporting outcomes dishonestly might violate participant's moral values, and/or undermine the appearance of being honest. Hence, participants face a trade-off between the financial gain from reporting a head and the psychological cost of misreporting a head. The psychological cost of misreporting might be greater when participant reports through a human voice channel as compared to a voicebot. Hearing a human voice and anticipating reporting to an experimenter can heighten the salience of unethical behavior, potentially prompting individuals to avoid engaging in actions that would signal an undesirable identity. This is because people are motivated to maintain a positive self-image, and avoid behaviors that could contradict it.[14]

[12] Like us, Biener and Waeber (2021) developed a dedicated chatbot interface in lieu of the Skype machine recording to resolve the problems with lack of anonymity and uniformity.

[13] The instructions of the entire experiment can be found online via this link: https://www.dropbox.com/s/bigt227oxbp7g69/chatbotsurvey.pdf?dl=0

[14] Bénabou and Tirole (2006, 2011) developed a theoretical framework based on a cognitive model of identity management. This model posits that when choosing an action, individuals take into account what kind of person the choice would make them and how desirable achieving a self-image is to them. Recent experiments provide support for this phenomenon; for example, Falk (2021) show that an increase in the salience of self-image through the use of mirrors reduces the incidence of immoral behavior and participants were less likely to inflict a painful electric shock on others to receive a monetary reward. Capra et al. (2021) show that altruistic self-concept can mediate ethical behaviors.

The experimental literature on the effect of reporting through a human voice vs. a bot voice channel on misreporting is scarce. The pioneering paper by Cohn et al. (2022) is closest to ours, but its main objective is different. Their paper aims at studying cheating with and without *direct human interaction*. In contrast, we are interested in isolating the influence that the reporting channel has on misreporting. Cohn et al. (2022) found that reporting to a machine increases the propensity of lying as compared to reporting to a human. Another related experimental study by Biener and Waeber (2021) compared behaviors when reporting to a chatbot vs. a human. The authors found that, when a person interacts with another human who's agency is made salient, social-image concerns reduce dishonesty. Clearly, the presence of another person may awaken self-image and social-image concerns that increase the psychological costs of misreporting, and these papers suggest that *human presence* is key to mitigating dishonest behavior. However, it is possible that self-image concerns may also be triggered when reporting through a human voice without a direct interaction with a human. Indeed, *f*MRI studies that investigate human-machine interactions have shown that the expectation of interacting with a human, not necessarily the human presence, activates social cognition networks in the brain that are not active when humans expect to encounter a machine (Ekins, 2013). This suggests that the expectation to report the coin-toss outcomes to the experimenter by hearing a prerecorded human voice may also trigger image concerns, resulting in a different response as compared to reporting to the experimenter via a voicebot. Based on these observations, we put forward the following hypothesis:

Hypothesis 1 (HumanVoice vs. VoiceBot)*: When reporting through a human voice channel as opposed to a bot channel, participants are less likely to report a head.*

In order to determine if dishonest behavior when interacting with chatbots could be reduced, Biener and Waeber (2021) varied the perceived agency of the chatbot by signaling sophistication via a variation in the interaction content. In their study, the sophistication was signaled by the following statement: *"I can process a lot of data and interactions at the same time, which is, I believe, the reason, I'm being involved in this survey. At the same time, I can mimic conversation-like interactions, which are closer to human communication than online forms."* It is possible that the interaction with a machine that has intelligence triggers social-image concerns (i.e., the person may feel observed and judged). However, the authors found that endowing the chatbot with an indication of sophistication did not induce more honest behavior. In our experiment, we also varied the sophistication of the voicebot. In contrast to Biener and Waeber (2021), we introduce our AIBot with the ability to detect emotion from voice analysis, which we believe can deepen the social-image concerns and thus reduce misreporting.

Hypothesis 2 (Bot Sophistication)*: Perceived higher voicebot sophistication (i.e., emotion detection ability) reduces the propensity to report a head.*

Although verbally reporting more heads than observed may raise concerns about one's morality, written lies may be perceived as more permanent and therefore may result in less dishonesty. Studies based on self-reports (Hancock et al., 2004; Markowitz, 2022) show that people tend to tell more lies through verbal communication as compared to written communication, this is because, unlike text, voice conversations are typically not recorded. However, in the coin toss task, Cohn et al. (2022) found that the likelihood of lying when reporting the results in written form through a web page link was the same as when using an automated voice response system (i.e., a prerecorded voice system). In light of this latter result, we expect no difference in lying behavior when reporting verbally or in written form.

Hypothesis 3 (Text vs. VoiceBot)*: There is no difference in the likelihood of reporting a head via text vs. via voice.*

in the voice treatments, the reportee's voice could be from a female or a male. Guo et al. (2020) demonstrate that female voice attributes have positive effects on male customers while male voice attributes have negative effects on both genders. Just like human voice, bot voice features provide social cues that influence the way humans interact with the bot (Feine et al., 2019); however, it is unclear whether male voice attributes such as trustworthiness or female voice attributes such as warmth have a stronger effect on misreporting.

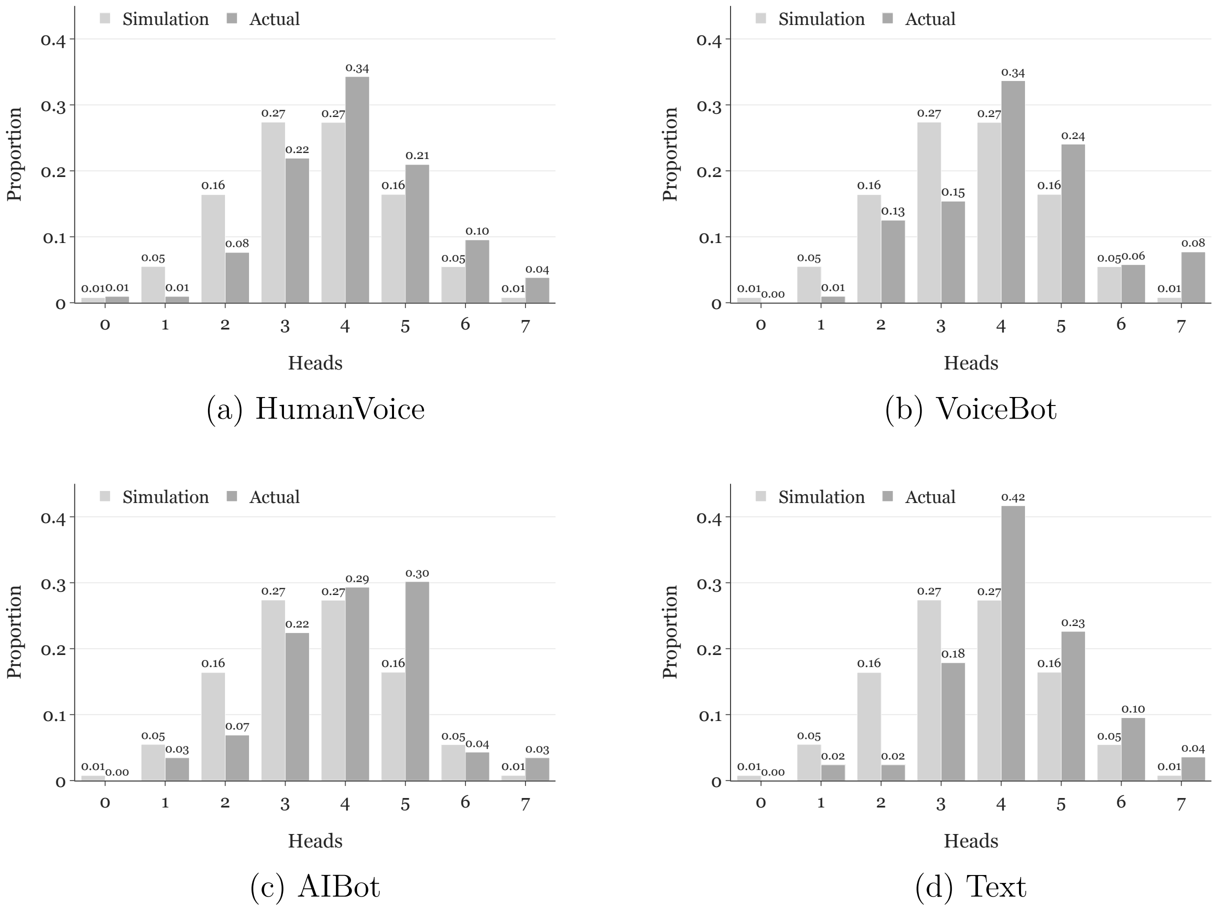
In our voice treatments, we vary the gender of the reportee's voice (female voice vs. male voice). Gender cue is one of the important and practical design features to equip voicebots (Feine et al., 2019). Guo et al. (2020) demonstrate that female voice attributes have positive effects on male customers while male voice attributes have negative effects on both genders. Just like human voice, bot voice features provide social cues that influence the way how humans interact with bots (Feine et al., 2019). The perceived attributes (i.e., warmth, attractiveness, etc.) of a feminine voice may have the effect of dissuading men to lie to a feminine voice. Likewise, the perceived attributes (i.e., authoritativeness, trustworthiness, etc.) of a masculine voice may also dissuade both men and women to a lie to a masculine voice. Thus, it's unclear which voice have a stronger effect on misreporting.

Hypothesis 4 (Feminine vs. Masculine Voice)*: The propensity to report a head is not affected by the gender of the reportee's voice.*

1.3 Results

Overall, our participants reported a head 58.7% of the time, which is significantly higher than the expected probability of tossing a head with a fair coin-flip (Binomial test: $p < 0.001$). In our experiment, the probability of misreporting was 17.4% which is within the range observed in Cohn et al. (2022) who found as little as 7% and as high as 24% misreporting depending on the experimental treatment.[15] Although their participants were students enrolled at a university whereas ours were recruited from Polific, a meta-analysis by Gerlach et al. (2019) shows little difference between lab and online behaviors in the rate of misreporting in coin toss tasks. Indeed, the cheating rate in our experiment ranged from 15% in the AIBot treatment to 20.8% in the Text treatment.

Figure 1: Distribution of Successful Coin Tosses (Actual vs. Simulation)



(a) HumanVoice

(b) VoiceBot

(c) AIBot

(d) Text

[15] Let h be the probability that a participant reports a head conditional on the actual coin toss being a tail. We assume that no one cheats to their disadvantage. If the coin toss is a tail, participants will report a head with a probability of h. If the outcome of a given coin toss is a head, participants will report a head with a probability of one. Thus, the unconditional probability of reporting a head p is given by $p = 0.5 * (1 + h)$. Based on the law of large numbers, we can then replace the probability of reporting a head with the average percentage of reported heads in a given condition to determine the cheating rate in that condition. Thus, the probability of misreporting a head is given by $h = 2 * p - 1$.

The four panels in Figure 1 contrast the distribution of the total number of reported heads and the expected distribution under truthful reporting by treatment. The expected distribution was obtained after 10,000 iterations of the coin toss task by the same sample size of each treatment. For example, in the VoiceBot treatment, 104 participants flipped a coin 7 times, so we simulated 728 such tosses and iterated the procedure 10,000 times. The top panels of Figure 1 show the distributions of expected (light colored bars) and reported (dark colored bars) heads in the HumanVoice and VoiceBot treatments, respectively. The bottom panels contrast the expected and observed distributions for the AIBot and Text treatments. Overall, the distributions of the observed number of heads reported are to the right of the simulated distributions for all treatments. This implies that participants over-reported the total number of observed heads. More specifically, participants under-reported 1, 2, and 3 heads and over-reported 4, 5, 6 and 7 heads.[16] Figures A1a - 1c in the Appendix show the simulated kernel distributions of heads and the observed average proportion of heads in each treatment.

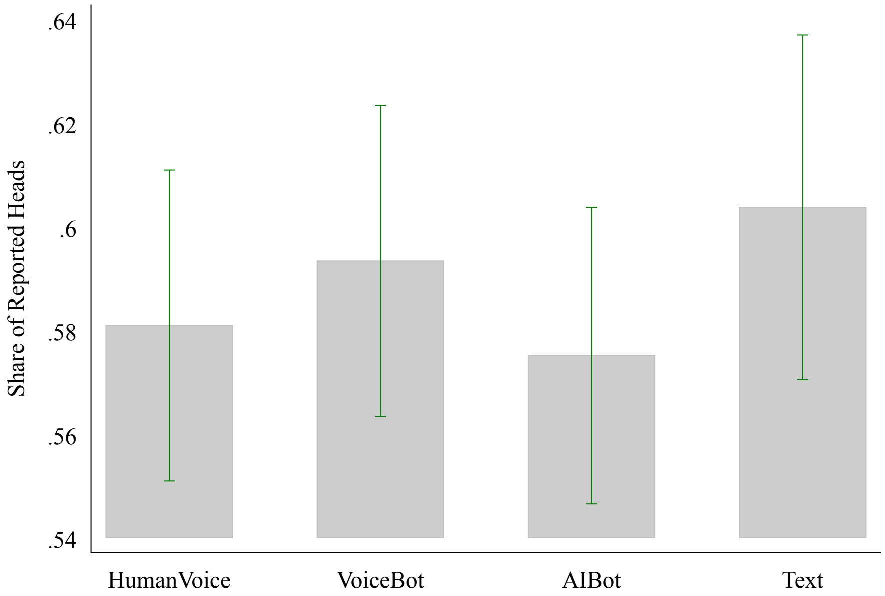
Result 1 (HumanVoice vs. VoiceBot): Our participants reported a head 58.1% of the time in the HumanVoice treatment and 59.3% in the VoiceBot treatment (see Figure 2), corresponding to a cheating rate of 16.2% and 18.6%, respectively. The differences in the proportion of reported heads are not statistically significant (Mann-Whitney; p=0.629).[17] This result indicates that participants are neither less nor more likely to report dishonestly through a human voice channel as opposed to a voicebot channel. Given that there is no difference between these two treatments, we are not surprised to see that when the voicebot is endowed with emotion detection abilities, the proportion of heads reported in the AIBot

[16] Only when comparing 6 heads in the AIBot treatment we observe a slightly lower proportion of heads reported compared to the simulation. However, our observations that depict extreme behaviors are too few to make any statistical inferences.

[17] The distributions of the total number of heads reported in these two treatments are not significantly different (p=0.85, Kolmogorov-Smirnov test).

and HumanVoice are also not significantly different (Mann Whitney; p=0.817).

Figure 2: Proportion of Reported Heads by Reporting Channels

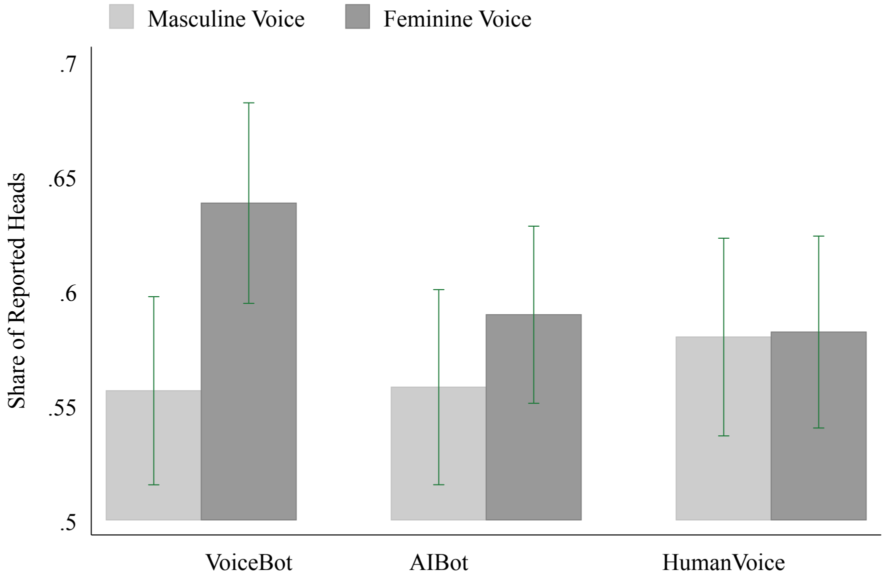


Result 2 (Bot Sophistication): The proportion of reported heads is 57.5% when reporting to an AIBot, which translates to a cheating rate of 15%. This does not significantly differ from the proportion of reported heads in the VoiceBot treatment (Mann Whitney; p=0.468) (see Figure 2), which implies that the propensity of misreporting is not affected by the sophistication (i.e., the ability to detect emotion) of the voicebot.

Result 3 (Text vs. Voice): When reporting via text, our participants reported a head 60.4% of the time which is slightly higher than reporting via voice (i.e., VoiceBot), 59.3% (see Figure 2). However, the difference is not statistically significant (Mann Whitney; p=0.704). This result tells us that there is no difference in the likelihood of reporting a head to a machine either via text vs. via voice.

Result 4 (Feminine vs. Masculine Voice): When we look at the aggregate results by the gender of the reportee's voice, we find that participants reported heads 60.4% of the time when they hear a feminine voice. This corresponds to a cheating rate of 20.8%. This rate is significantly higher than the rate of reported heads when hearing a masculine voice, 56.3%, corresponding to a cheating rate of 12.6% (Mann Whitney; p=0.042). Contrary to our Hypothesis 4, the propensity of reporting a head is affected by the gender of the reportee's voice. More specifically, feminine voice increases the propensity of misreporting as opposed to masculine voice. Decomposing the data by treatments, we find that the proportion of reported heads to a feminine voice is significantly higher when participants report to a VoiceBot (64% vs 56%, Mann Whitney; p=0.025), but there is no voice gender effect when reporting to a HumanVoice or AIBot (see Figure 3).

Figure 3: Proportion of Reported Heads to Feminine and Masculine Voice



1.3.1 Regression Results

In this section, we present the results of our logistic regression analysis. This analysis could help us identify the variables that influence the likelihood of reporting a head in each of our treatments. Table 1 - 4 show the average marginal effects of regressing a head flip on our treatments along with the participants' individual characteristics such as age, education, race and gender, attitudes toward chatbots and AI, trust and lying behaviors (i.e., levels of trust toward oneself and others, and self-reported frequency of lying), and attitudes toward feminine and masculine voices. As our survey included a large number of questions regarding participants' attitudes, we used factor analysis (Basilevsky, 2009) to reduce the dimensionality of the variables. Please refer to Section C in the Appendix A for details.

Table A3 - A6 in the Appendix show that for variables that measure: a) attitudes toward chatbots and AI, b) attitudes toward trust and lying behaviors, and c) attitudes toward feminine and masculine voices, we have three factors explaining at least 80% of the variability of those variables in each group. The three factors for attitudes toward chatbots and AI are: positive attitude toward chatbot (ChatbotPositive), negative attitude toward AI (AINegative), and familiarity with AI (AIFamiliar). The factors for trust and lying behaviors include: self-reported frequency of lying (Honesty), trust attitude toward other people (Trusting), and self-sense of trustworthiness and fairness (Trustworthiness). Our feminine and masculine voice attitudes have factor groupings which are related to: attractiveness and familiarity (Attractive), trustworthiness and likeability (Trustworthy), and authoritativeness and confidence (Authoritative). We used the above factors in our regression analyses when indicated.

In Table 1 Columns (1) and (2), we present the regression results with HumanVoice as a baseline. The estimated coefficients in Column (1) reveal that reporting via Text, reporting to a VoiceBot, or reporting to an AIBot has no influence on the propensity of

Table 1: Treatments and Variables That Influence Reporting Heads

	(1)	(2)	(3)	(4)
HumanVoice			-0.011	-0.007
			(0.025)	(0.027)
VoiceBot	0.013	0.011		0.003
	(0.026)	(0.025)		(0.026)
AIbot	-0.006	0.000	-0.011	-0.008
	(0.024)	(0.023)	(0.023)	(0.024)
Text	0.023	0.007	-0.003	
	(0.026)	(0.027)	(0.026)	
Female		-0.088***	-0.088***	-0.088***
		(0.018)	(0.018)	(0.018)
College		0.006	0.006	0.006
		(0.018)	(0.018)	(0.018)
White		-0.016	-0.016	-0.016
		(0.021)	(0.021)	(0.021)
Age		-0.002**	-0.002**	-0.002**
		(0.001)	(0.001)	(0.001)
Honesty		-0.007*	-0.007*	-0.007*
		(0.005)	(0.005)	(0.005)
Trusting		-0.013**	-0.013**	-0.013**
		(0.007)	(0.007)	(0.007)
Trustworthiness		-0.018**	-0.018**	-0.018**
		(0.008)	(0.008)	(0.008)
Observations	2863	2821	2821	2821
Subjects	409	403	403	403
Controls				
Chatbot & AI Attitudes	No	Yes	Yes	Yes

Column (1) and (2) show the marginal effects using HumanVoice as the base category, while Column (3) uses VoiceBot and Column (4) uses Text as base category, respectively. Age is categorical variable. Clustered standard errors are in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

reporting a head compared to reporting to a human voice. This implies that, in general, the mode of reporting has no effect on the tendency to misreport, confirming what we previously observed in **Results 1 - 3**. Adding demographic information and other controls, including attitude toward chatbots and AI, level of trust to other (Trusting), self-sense of trustworthiness (Trustworthiness), and self-reported frequency of lying (Honesty), does not change the results as shown in Column (2). In Columns (3) and (4), our baseline variables are VoiceBot and Text, respectively. Column (3) shows that the sophistication of the bot does not change the reporting behavior. Column (4) shows there is no effect of verbally reporting in comparison to just using text. This implies that reporting to a prerecorded voice (hearing a voice and verbally reporting heads) does not reduce the propensity to misreport compared to reporting via text.

With respect to the determinants of misreporting we find that, although the propensity of reporting a head is similar in all the treatments, there are significant gender differences in misreporting (see Columns (2) - (4) in Table 1). The coefficient of Female indicates that female participants are less likely to report heads by 8.8% (p<0.000) compared to male participants. This result is in line with the findings of Dreber and Johannesson (2008a) and Guerra et al. (2022). Furthermore, we find that when reporting via Text, VoiceBot or AIBot, female participants are less likely to misreport than male participants by 8.6% (p=0.013), 15.1% (p<0.000) and 11.6% (p=0.001), respectively. We do not observe differences in the proportion of reported heads in the HumanVoice treament (please refer to Table A7 in the Appendix). Moreover, older age cohorts are less likely to report a head, which is consistent with previous studies. In an experiment with children, Zhang et al. (2020) found that older cohorts tend to be more honest. Interestingly, attitudes toward lying and trust do influence the propensity to misreport. More specifically, the more trusting and trustworthy individuals are, the less likely they are to report a head.[18]

[18] This is in line with the finding by Markowitz (2022), which suggests that individual characteristics explain more of the variance in lying than the communication medium (e.g., face-to-face communication,

Table 2: Reporting Heads to Feminine/Masculine Voice within Treatment

	(1) **VoiceBot**	(2) **AIbot**	(3) **HumanVoice**
Feminine Voice	0.070*	0.035	-0.020
	(0.036)	(0.032)	(0.034)
Subjects	102	115	103
Observations	714	805	721

The table shows the average marginal effects. Demographic and attitudinal variables (i.e., chatbot and AI, trust and lying) are controlled in the regression. Clustered standard errors are in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

In our experiment, participants were randomly assigned to a reporting channel that had a feminine voice or a masculine voice. Table 2 shows the results regressing reporting a head on feminine voice within each reporting channel treatment controling demographic and attitudinal variables. Column (1) shows that participants were 7% (p=0.053) more likely to report a head to a female VoiceBot as opposed to a male VoiceBot. We did not observe any voice gender effects in HumanVoice (Column (3)) or AIBot (Column (2)) treatments. These results confirm what we previously observed in **Result 4** in which the propensity of reporting a head is affected by the gender of the reportee's but only in VoiceBot treatment.

To compare and contrast the voice gender effects across different voice channels, we regressed a head flip on the voice gender (Feminine Voice) with the HumanVoice treatment as our baseline in Table 3. Column (1) shows the initial results. The propensity to report a head is 3.6% (p=0.074) higher when participants report to a feminine voice. To test for differences in reporting a head across treatments when participants hear a feminine voice, we added interactions (see Column (2)). There is a marginally significant higher likelihood (6%) of reporting a head when participants hear a feminine voice in the VoiceBot treatment compared to the HumanVoice treatment (VoiceBot on feminine voice, p=0.099). Meanwhile, as shown in Column (2), when reporting to a masculine voice, the propensity of reporting

social media, texting, the phone, video chat, and email.).

Table 3: Reporting Heads to Feminine/Masculine Voicebot by Treatment

	(1)	(2)
HumanVoice		
Voicebot	0.022	-0.0122
	(0.025)	(0.035)
AIbot	0.002	-0.008
	(0.025)	(0.034)
Feminine Voice	0.036*	0.005
	(0.020)	(0.036)
Voicebot x Feminine Voice		0.072
		(0.051)
AIbot x Feminine Voice		0.023
		(0.047)
Interaction Effects		
Voicebot on feminine voice		0.060*
		(0.036)
AIbot on feminine voice		0.014
		(0.034)
Subjects	320	320
Observations	2,240	2,240

The table shows the average marginal effects using HumanVoice as the base category. Demographic and all attitudinal variables are controlled in the regression. Clustered standard errors are in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

a head is not statistically different between the two treatments (please refer to VoiceBot coefficient in Column (2)). This indicates that voicebots equipped with a masculine voice could deter misreporting and make people behave closer to how they would behave when facing a human. Thus, our findings suggest that when replacing humans with voicebots, it is suggested to use a masculine voice to deter misreporting.

Furthermore, our analysis shows that there is no significant difference in the likelihood of reporting a head between the AIBot and HumanVoice treatments, regardless of whether the participant was reporting to a masculine or feminine voice (see AIBot and AIbot on feminine voice). These results imply that the sophistication of the voicebot could reduce the voice gender effect that we previously observed in the VoiceBot treatment (see Table 2). Thus, when replacing humans with more sophisticated voicebots, either feminine or masculine voice could be utilized.

Table 4: Reporting Heads to Feminine/Masculine Voicebot by Participants Gender

	(1) **Feminine Voice**	(2) **Masculine Voice**
Human Voice		
Voicebot	0.177***	0.029
	(0.062)	(0.053)
AIbot	0.065	0.0169
	(0.063)	(0.053)
Female	-0.054	0.0234
	(0.057)	(0.057)
Voicebot x Female	-0.192**	-0.108
	(0.080)	(0.073)
AIbot x Female	-0.052	-0.081
	(0.074)	(0.067)
Interaction Effects		
Voicebot on female participants	0.015	-0.078*
	(0.043)	(0.046)
AIbot on female participants	0.013	0.064
	(0.041)	(0.042)
Subjects	164	156
Observations	1,148	1,092

The table shows the average marginal effects using HumanVoice as the base category. Demographic and all attitudinal variables are controlled in the regression. Clustered standard errors are in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

In Table 1 we found that female participants were, on average, 8.8% less likely to report a head. Previous study shows that feminine voice attributes have positive effects on male customers while masculine voice attributes have negative effects on both genders (Guo et al., 2020). In the context of misreporting, we are interested in finding out how our participants behave when reporting to a feminine voice as compared to a masculine voice. Table 4 presents the results of the regressions. In Column (1) and (2), we present the regression for feminine voice and masculine voice, respectively, with our baseline treatment being HumanVoice. Column (1) shows that male participants were 17.7% (p=0.004) more likely to report a head when reporting to a female VoiceBot compared to reporting to a female HumanVoice. Although, in general, the reporting channel does not have direct effect on misreporting, when interacting with feminine voice, it has big impact specially on male participants. Interestingly, male participants do not change their reporting behavior across reporting channels when hearing a masculine voice (Column (2)). The interaction effects suggest that there is no significant different in misreporting across reporting channel treatments for female participants when hearing a feminine voice. However, we found that when reporting to a masculine voice, female participants in the VoiceBot treatment were roughly 7.8% less likely to report heads than in HumanVoice treatment, although the effect is small and only marginally significant (p=0.088). These results suggest that using a masculine voice is advised when replacing human with a voicebot, as it does not induce more dishonesty in reporting behavior for either the male or female users. However, in specific case when feminine voice is a required feature, then it is advised to utilize a more sophisticated voicebot that is able to analyze voice, as male users tend to report more dishonestly when hearing a feminine voicebot.

All in all, the propensity to report a head via a human voice channel is not affected by the gender of the participant nor by the gender of the reportee's voice. However, both variables influence missreporting behavior when reporting to a voicebot. Interestingly, with the increase in the level of sophistication of bot sophistication (i.e., AIBot), the voice gender effect is no longer significant indicating that the sophistication of the voicebot could reduce the gap in misreporting behavior in human-machine interactions.

1.4 Discussion and Conclusion

This paper contributes to a small, but growing body of experimental studies on human-machine interactions. Organizations are becoming increasingly reliant on chatbots, voicebots, and similar AI-driven technology to provide customer services and support. Similarly, governments are adopting AI technology and online platforms for e-procurement systems. However, it is still unknown whether what we have learned from studies on deception and misreporting in human-human interactions extend to situations where humans interact with machines. Should organizations expect more dishonest behaviors with the transition to AI-driven customer service? Similarly, should governments expect more misreporting? It is important to answer these questions because dishonest reporting can not only generate direct financial costs, but it can also compromise the accuracy of data fed into an organization's systems, leading to poor decision-making. For instance, chatbots and voicebots that collect data and feedback from customers can produce inaccurate insights and incorrect assumptions about customer behavior, preferences, and needs if customers provide false information.

We designed an experiment to study dishonesty in reporting through bot vs. human voice channels. In our experiment, we utilized the coin-toss task (Bucciol and Piovesan, 2011a; Houser et al., 2012b) to measure the propensity to misreport. The task involves rewarding participants for reporting more heads in a series of coin tosses and only the reported outcomes are observable to the experimenter. We focus on the effect of different reporting channels on dishonest reporting. Specifically, we investigate whether people are more likely to report heads when they report through a voicebot vs. a human voice. By isolating the reporting channel, we eliminate the effect of a human presence on dishhonest behavior. We recruited participants through Prolific, who were older and wealthier than most student subjects recruited at university settings. Our participants were randomly assigned to one of four treatments that differed with respect to the kind of reporting channel: reporting via a HumanVoice channel, via a VoiceBot channel, via an AI enhanced voicebot channel

(AIBot), and via Text.

Contrary to a similar study by Cohn et al. (2022) who showed that reporting to a machine increases the propensity of lying compared to reporting to a human, we found no differences in misreporting across HumanVoice, AI-enhanced VoiceBot, and VoiceBot reporting channels. There are several reasons for why our results differ. To begin, while their paper focuses on the effect of human-machine interactions, our paper focuses on comparing dishonest reporting through human voice vs. machine reporting channels; that is, we abstract from the effect of a human presence. Furthermore, unlike our paper, participants in Cohn et al. (2022) interacted with the experimenter first and, although dishonesty by a particular participant was impossible to identify, their reporting mechanism did not ensure full anonymity.[19] To isolate the effect of reporting channels, we developed a dedicated online interface for the voicebot and the human voice, allowing the participant to enter her reports after being prompted by a pre-recorded voice without having to interact with an experimenter.[20] Our results are closer to the results by Biener and Waeber (2021) who found no difference in the propensity of misreporting when comparing a generic chatbot to a generic human, or comparing a generic chatbot to a "sophisticated" chatbot, or comparing a generic human to a sophisticated chatbot. There is difference only when comparing a generic chatbot to a "sophisticated" human. In their study, the sophistication of the human was signaled by the following statement: *"I'm a talented student, which is, I believe, the reason, I'm being involved in this survey. I also have good communication skills, helping me to understand others, and to be understood."* [21]

[19] Human reporting was implemented by asking participants to call or text the experimenter and report directly to her; that is, there was interaction between the experimenter and the participant in these treatments. Machine reporting was implemented via Skype and an answering machine in which usernames, profile picture, or phone numbers could be made available to the experimenter. Text was implemented via a weblink.

[20] This link takes you to the sample interface: https://aixecon.com/chatbot_F (please use Google Chrome or Safari to access, type "999" as your ID#, and allow microphone access. When encountering a connection issue, please click advanced option and continue to "proceed to the website". This will allow the connection to access the website.).

[21] In Biener and Waeber (2021), the human treatment was introduced by the following statement: *"Hello!*

Taking into consideration the variations in experimental designs between our study and prior research, we can reconcile the inconsistent findings with prior research by concluding that the expectation of reporting outcomes, whether through a human voice channel, a voicebot channel, or an AI-enhanced voicebot channel, does not significantly influence dishonest behavior. However, the presence of an actual human being seems to play a key role in nudging individuals toward greater honesty. In other words, the mere presence of a human seems to act as a "truth serum".

We also found that compared to using text to report the coin-toss outcomes, reporting via voice did not increase misreporting. This result contradicts studies based on self-reports (Hancock et al., 2004; Markowitz, 2022) where the authors showed that people tended to tell more lies through verbal communication as compared to written communication; unlike text, voice conversations are typically not recorded. However, in our experiment, participants were made aware that both text and voice reports were recorded. Similarly, in Cohn et al. (2022) the authors found that the likelihood of lying when reporting the results in written form through a web page link was the same as when using an automated voice response system (i.e., a prerecorded voice system).

In our experiment, we use voice as the main channel for reporting coin tosses which can either be a feminine voice or a masculine voice. We find that reporting to a feminine voice induces more dishonesty compared to reporting to a masculine voice (60.4% vs. 56.3%, p=0.042). The difference is significant when reporting through a voicebot channel and disap-

I'm a student assistant, helping the experimenters to collect your responses for this survey. Are you ready?", the generic chatbot was introduced by the following statement: *"Hello! I'm an automated chatbot, the experimenters programmed to collect your responses for this survey. Are you ready?"*, the sophistication of the chatbot was signaled by the following statement: *"I can process a lot of data and interactions at the same time, which is, I believe, the reason, I'm being involved in this survey. At the same time, I can mimic conversation-like interactions, which are closer to human communication than online forms."* In our experiment, we also varied the sophistication of the VoiceBot. However, we introduced our AIBot with the ability to detect emotion from voice analysis, which we believe can heighten the sense of observability and thus affect misreporting, by the following statement: *"Emile, an artificial intelligence (AI) driven emotion detection chatbot that analyzes voice speed, intonation, and tone."*

pears when reporting via a human voice channel. These results suggest that companies could utilize male voicebots to deter or diminish dishonest behaviors. Recall that participants did not know the gender of Emile until they heard her\him talk through our interactive platform. The fact that the likelihood of reporting heads was higher when our participants reported through a female voicebot channel suggests that *hearing* the feminine voice triggered them to be less honest. In a large study on personality impressions based on voice, McAleer et al. (2014) showed that feminine voice was associated with traits such as friendliness, likeability and warmth whereas masculine voice is associated with dominance and competence. Perhaps people find it easier to lie to a feminine voice because it is perceived as more forgiving. Interestingly, when we disentangle the voice gender effect by the gender of participants, we find that the higher likelihood of reporting a head to a female voicebot is driven by male participants.

Regarding gender differences in misreporting, we find evidence of women lying less than men by 8.8% (p<0.000). However, when reporting through the human voice channel, no differences between men and women were observed. Notably, male participants were more affected by our experimental treatments than female participants. Men's likelihood of misreporting decreased with the perceived sophistication of the voicebot channel. Women, on the other hand, were more immune to the kind of channel they reported through. Although the mechanisms that cause the difference are unknown, perhaps the difference in behaviors between men and women may have stemmed from gender-related differences in moral reasoning. Lohse and Qari (2021) showed that men lied significantly less when there was a chance to be audited whereas women's lying behavior was unchanged. Similarly, Dreber and Johannesson (2008a) and García et al. (2021) showed that men were more prone to behave dishonestly than women when there was a reward, whereas there was no gender difference in dishonest behavior when there is no reward. Cordellieri et al. (2020) showed that women were less prone than men to accept a moral violation. Friesdorf et al. (2015) found that men showed a stronger preference for utilitarian over deontological judgments while women exhibited stronger deontological inclinations.[22]

In our study, we link lying with moral violations and honesty with sacred values which affect behavior through deontic rules and not through a narrow utilitarian evaluation of costs and benefits (Berns et al., 2012). As such, while female participants are neither more nor less likely to misreport their coin toss outcomes when reporting via a human voice than when reporting via a voicebot, male participants adjust and change their behaviors. When reporting to a human voice or a sophisticated bot, the psychological cost of misreporting might be greater as participants may feel observed or perhaps even judged.

Interestingly, we also found that the sophistication of voicebot could reduce the differences in dishonest reporting not only between reporting through a feminine voice and a masculine voice, but also between women and men. More specifically, the increase in misreporting when reporting to a feminine voice disappeared in the AIBot treatment and the effect of participants' gender became smaller when reporting to an AIBot. In other words, AI-enhanced voicebots could make individuals behave in a way that is closer to how they behave when interacting with a human. For organizations interested in deterring or diminishing dishonest reporting in human-machine interactions, these results suggest that developers may want to consider utilizing a male voicebot or an AI-enhanced, sophisticated female voicebot.

All in all, this paper contributes to the existing literature on misreporting in human-machine interactions. Per our knowledge, this is the first paper that evaluates reporting channels without adding a human presence. Our findings can inform voicebot development in business application, especially in how feminine and masculine voice cues influence female and male users or reporters. Lastly, our results complement the existing literature about the development of human-machine interactions especially in how human behavior toward

[22] The principle of deontology states that the morality of an action depends on its consistency with moral norms; the principle of utilitarianism implies that the morality of an action depends on its consequences (Friesdorf et al., 2015).

machine may change with the increasing level of bot sophistication.

Moving forward, this paper can motivate interesting follow-up studies. Could the interaction with a sophisticated bot make a human being more honest and conscientious thereby reducing the tendency to engage in immoral or fraudulent behaviours? In the 1960s, MIT developed a pioneering artificial intelligence program called ELIZA. ELIZA was designed to engage in conversations with people using scripted responses that mimicked a psychotherapist's dialogue. While ELIZA was not capable of true understanding or empathy, it was surprisingly effective in getting people to open up and share personal information with it. That is, people became in a way "more human" when interacting with ELIZA. As technology advances and machines become more sophisticated, it becomes increasingly important to understand the psychological and social effects of these interactions and their influence on human behaviors. In particular, given the enormous costs that are associated with fraudulent reporting, machines could be designed to have features that reduce misreporting saving organizations billions in lost revenues. Another interesting next step in this line of research is to examine the effects of back-and-forth interactions between humans and voicebots. For example, how do people respond differently to a machine that is designed to mimic human-like conversational norms versus one that is purely functional? These are crucial questions that can help us understand the nature of human-machine interaction and its impact on human behaviors.

Chapter 2

The Sound of Cooperation and Deception in High-Stake Events

Coauthored with C. Mónica Capra and Matthew Gomies

2.1 Introduction

Communication is ubiquitous in all human strategic interactions. An extensive body of game theoretic and experimental research has been devoted to understanding its role in influencing behavior. Although communication in the form of informal talk is thought to be uninformative when the interests of the players are not aligned (Farrell and Rabin, 1996), recent experiments in economics reveal that this form of communication is more than just cheap talk. It enhances the feelings of empathy by reducing social distance (Andreoni and Rao, 2011; Preston and De Waal, 2002; Bohnet and Frey, 1999), awakens feelings of guilt (Charness and Dufwenberg, 2006; Battigalli and Dufwenberg, 2007) and arouses the need to maintain a positive self-image (Bodner and Prelec, 2003; Bénabou and Tirole, 2006). These psychological mechanisms affect cooperative behavior. However, when the stakes are high and the interest of the players are not aligned, lies and deception do emerge in spite of one's intrinsic aversion toward dishonesty (Gneezy, 2005; Lundquist et al., 2009; Serra-Garcia et al., 2011; Erat and Gneezy, 2012a; Cappelen et al., 2013; Abeler et al., 2014).

Linguists rely on scripted verbal statements and written messages to identify cues of deception (Larcker and Zakolyukina, 2012; DePaulo et al., 2003; Hauch et al., 2012). Scripted conversations allow researchers to extract bag of words and categorize statements based on language structure. In game experiments, Capra (2019) and Penczynski (2019) used text messages and scripted conversations to identify player types and emotions. To identify intentions, Turmunkh et al. (2019) analyzed scripted conversations from a high-stakes TV game show that mimics a prisoner's dilemma game. The authors show that conditional or implicit statements reveal intention not to cooperate. Yet, despite the usefulness of studying text, much information about personality, emotions, intentions, and attitudes is contained in the non-verbal aspects of communication like voice, facial expressions, and body posture (Mehrabian and Wiener, 1967; Mehrabian and Ferris, 1967; Afifi, 2007; Mandal, 2014; Phutela, 2015).

In this paper we investigate whether acoustic features of the voice predict cooperative and deceptive behaviors in real high-stakes interactions. Voice is as unique as one's finger print; it provides cues about a person's sex, age, and body size. Studies have shown that voice can convey information that may directly influence decision making (Hughes et al., 2009; Hughes and Rhodes, 2010; Pisanski et al., 2014; Lassalle et al., 2019; Allport and Cantril, 1934). For example, McAleer et al. (2014) and O'Connor and Barclay (2017), show that acoustic features such as pitch and voice roughness are linked to perceived trustworthiness, reliability, competence, and likeability. In contrast to verbal statements, voice modulation is in large part non-volitional and spontaneous (Lavan et al., 2019), and although people can hide their intent to deceive by manipulating their language, their voice will still capture the psychological cost of lying (Sen et al., 2020; Ekici et al., 2017; Sondhi et al., 2016). Thus, we believe that voice may be more likely than written text to capture much of the subconscious psychological processes whereby communication affects cooperative and deceptive behaviours in strategic interactions.

To investigate the predictive power of voice, we analyze the recordings from contestants' conversations in the British TV show "Golden Balls" and from recorded oral depositions in "Real-life Trial". In the former two contestants engage in face-to-face pre-play communication before deciding whether to split or steal a large jackpot. The game show mimics a prisoner's dilemma game with split representing cooperation and steal representing defection. Like Turmunkh et al. (2019) we are interested in identifying intent to cooperate (i.e., choose to split the jackpot), but unlike these authors, we concentrate on voice only, not scripted communication. Real-life Trial is a dataset collected by the University of Michigan AI lab that contains defendants' or witnesses' testimonies in real civil court proceedings (Pérez-Rosas et al., 2015). In Real-life Trial, it is possible to establish whether the recorded statements are deceptive based on the police investigation and the final verdict. For our analysis we extract voice and basic acoustic features from these two video sources, along with emotions from the voice and sentiments from the text.

For prediction we utilize decision trees and random forest. These two supervised machine learning methods are effective classifiers for small samples. While decision trees are easy to interpret and allow us to visualize features, random forest has higher prediction accuracy.[1] Our machine learning analysis shows that acoustic features are effective at predicting both cooperation and deception. Compared with the benchmark accuracy of naïve guess (i.e., 50%), we find that acoustic features for Real-life Trial have a significant improvement in prediction accuracy. The average prediction accuracy is 76% for females and 79% for males. Our predictions for Golden Balls achieve an average accuracy of 59% for females and 58% for males. The lower accuracy for Golden Balls as compared to Real-life Trial is partly due to the fact that contestants' conversations in the TV show were recorded over background music. Surprisingly, adding sentiment extracted from text does not significantly improve our prediction accuracy. We also find that specific features of the voice are more predictive of choices than others. Higher pitch, in particular, is associated with higher probability of deception in females. The results of males' voices are mixed. While high pitch is associated with higher probability of deception in Real-life Trial, it is also positively associated with

[1] Decision trees can be unstable as small variations in the data can result in large changes in the tree diagram. Decision trees are also prone to over-fitting. Random forest, a combination of multiple decision trees, can reduce the risk of over-fitting and produce more accurate results.

cooperation in Golden Balls. We also find that intonation (i.e., the standard deviation of pitch) is negatively associated with deception for both female and male in Real-life Trial. Those results are robust after controlling for sentiment from the scripted conversations. All in all, our results show that intent to cooperate and deceive in real high-stakes interactions is, at a minimum, partly captured by voice features alone.

Unlike text analysis that is usually context dependent, voice could allow us to see whether features predictive of deceptive behavior in one context are also predictive in a different context. In our samples, the defendants in Real-life Trial talk about their cases and can lie about what they have done/seen, whereas the contestants in Golden Balls talk about the game and can lie about what they are going to do (to split or to steal). Transferability is not possible to determine using text only, because language is completely specific to context. We test whether acoustic features of deception in the Real-life Trial can predict deception in Golden Balls, and vice-versa. However, our machine learning models show a very small improvement in prediction accuracy compared to a naïve guess, and only for females. This finding suggests that features of voice linked to deception in one context don't necessarily transfer to another context.

This paper proceeds as follows. Section 2 describes the data in more detail. Section 3 presents the methods that we use in our analysis. Section 4 provides the descriptive statistics of our data and the analysis results. Section 5 contains the discussion and conclusion.

2.2 Data

We collected voice data from two real-life high stakes video clip sources. The first source is "Golden Balls", a popular British TV game show which contains conversations of two people immediately before making decisions to cooperate or to defect in a prisoner's dilemma game. The second source is "Real-life Trial", which contains defendants' or witnesses' testimonies in civil court proceedings.

2.2.1 Golden Balls

Golden Balls (GB) is a British television game show broadcast on the ITV network between June 2007 and December 2009. In this show, two contestants interact in several rounds of games and collect money.[2] In the final round, the contestants play a game that resembles a classic prisoner's dilemma game with pre-play communication to decide how to allocate the money previously collected. More specifically, in the final round, each contestant is presented with two golden balls, one with the word "Split" written on it, and the other with the word "Steal" written on it. Contestants are asked to choose a ball at the same time. If both choose the "Split" ball, they split the jackpot equally. If both choose the "Steal" ball, both go home with nothing. If one chooses to split while the other chooses to steal, the one who steals receives the entire jackpot and the one who splits goes home with nothing. In this game, stealing weakly dominates splitting. Before deciding which ball to choose, the contestants are given a short time (between 1 and 5 minutes) to converse with each other. During this time, they can try to persuade the other player to cooperate (i.e., to choose split). Once the conversation ends, the contestants simultaneously choose their desired golden ball. During the 2 years of Golden Balls broadcast, there were a total of 323 episodes produced of which 118 are publicly available via YouTube. Each of these public episodes includes video recordings from two players for a total of 236 clips.

2.2.2 Real-life Trial

Real-life Trial (RT) is an online dataset that consists of 121 video clips from real court trials and video interviews. The dataset was created from defendants' trial recordings and video interviews where 61 deceptive and 60 truthful testimonies could be fairly observed

[2] In this TV show, 4 contestants will interact in several rounds of games to collect money. At the end of each round the contestants will vote to decide which contestant that should be eliminated. In the final round, there would be 2 contestants left to play a game to decide how to allocate the money previously collected.

and verified (Pérez-Rosas et al., 2015).[3] The determination of deception or truthfulness of a testimony is based on police investigations and the final verdict. For guilty verdicts, deceptive accounts were collected from a defendant in a trial and truthful testimonies were collected from witnesses in the same trial. In some instances, deceptive accounts were also collected from a suspect denying a crime he/she committed and truthful videos were taken from the same suspect when answering questions concerning other facts that were later verified by the police as truthful. For the witnesses, testimonies that were verified by police investigations were labeled as truthful whereas testimonies in favor of a guilty suspect were labeled as deceptive. Exoneration testimonies were obtained from "The Innocence Project" website and were labeled as truthful statements.[4] For our study, we selected 108 video clips with good audio quality. That is, of the original dataset of 121 videos, we excluded 13 poor quality clips, 7 of which were truthful videos.

2.3 Methods

In this paper, we investigate whether acoustic features of the voice during pre-play conversations in the GB game and oral depositions in RT can predict cooperative and deceptive behaviors. In GB, we analyze the contestants' voice during conversation before splitting or stealing a jackpot. Their conversations include neutral expressions and expressions of willingness, intention, and commitment to split. We then observe the contestants' final choice (split or steal). The conversation of contestants who chose to split was labeled as cooperative, the conversation of those who chose to steal was labeled as deceptive. About 56% of the contestants or a total of 132 people provided cooperative statements, 74 of those were females. In RT, we analyzed the testimonies in a court by defendants or witnesses that were

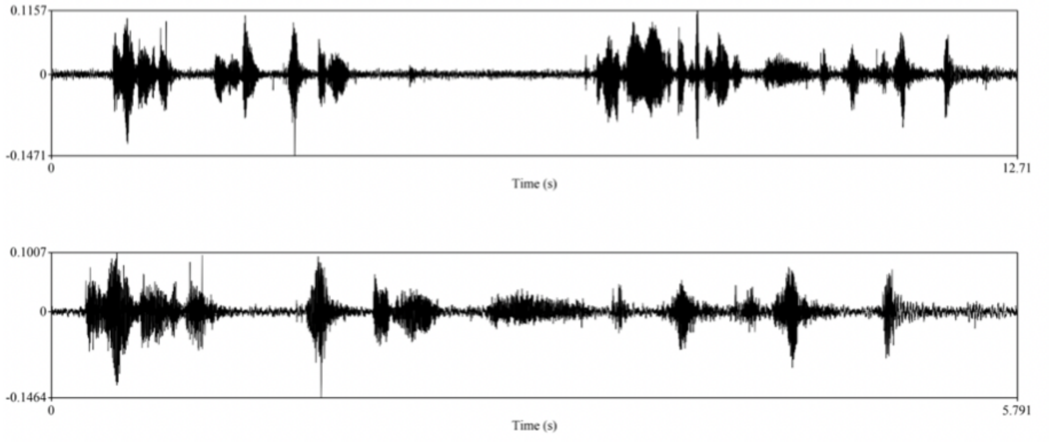
[3] Data are available and can be downloaded from: http://web.eecs.umich.edu/~mihalcea/downloads.html.

[4] Examples of famous trials included in the dataset are the trials of Jodi Arias, Donna Scrivo, Jamie Hood, and many others. For instance, Jodi Arias was charged with murdering her ex-boyfriend Travis Alexander in June 2008. The video clips contains her testimonies in the trial. The early statements that deny the murder were labeled as deceptive. The later statements that admit the crime were labeled as truthful.

later verified as truthful or deceptive statements. About 49% (53 out of 108) told the truth, and 21 of these were female defendants.

For both GB and RT, we first extracted audio clips from the video recordings via a multi-platform audio editor Audicity®.[5] We used this editor to convert all the collected video recordings into audio clips with the frequency value of 16000 Hz, bit rate value of 128 kbps, and in mono channel. We then isolated the speaker's voice by removing background noise and all the voices that are not originated from the main speaker (i.e., the host's voice in GB, the judges', attorneys' and prosecutors' voice in RT) and also normalized the volume using the same application. Figure 1 shows the voice waveform in two dimensional planes representation (normalized hertz and time in mili-second) before and after noise reduction and voice isolation performed using Audicity®. Once we isolated and normalized the speaker's voice, we extracted acoustic features such as pitch and intonation. We used a python library called Parselmouth for this purpose. Figure 3 depicts the process of data extraction. Section 2.3.1 describes these methods in details.

Figure 1: Voice Waveform Before and After Noise Reduction and Vocal Isolation



Note: The top figure shows an example of voice wave from RT before noise reduction and vocal isolation; the bottom figure shows the voice wave after noise reduction and vocal isolation.

[5] Audacity is a free and open-source digital audio editor and recording application software.

2.3.1 Features

In this section we describe the sets of features extracted from voice clips used to build classifiers of cooperation and deception. While our focus is on acoustic features of the voice, we also extract additional features such as emotion from voice, and sentiment from word transcript in order to evaluate the efficiency of acoustic features in prediction.

Acoustic Features

Following McAleer et al. (2014) and several past studies within the field of deception analysis (Sondhi et al., 2016; Sen et al., 2020), we employ the basic features of voice that are important in identifying social traits. These features include the fundamental frequency (pitch)[6], the standard deviation of the fundamental frequency (intonation), harmonic-to-noise ratio (HNR), voice perturbation measures like shimmer and jitter, and formants characteristics that represent resonant peaks in the spectrum. Pitch and intonation, in particular, are our main focus of analysis for this paper as they are an integral part of the human voice and have been shown to be indicative of psychological phenomena such as stress and lying (Sondhi et al., 2016). The pitch of the voice is defined as the rate of vibration of the vocal folds. The sound of the voice changes as the rate of vibrations varies. As the number of vibrations per second increases, so does the pitch, meaning the voice would sound higher. The human ear can detect a wide range of frequencies ranges from 20 to 20,000 Hz (audible to human ear). The pitch and its standard deviation (intonation) are unique to the individual, robust to other sonic influences, and require intensive training to modify them intentionally. The HNR is a measure of roughness of the voice and evaluates the efficiency of

[6] The fundamental frequency or F0 is the frequency at which vocal chords vibrate in voiced sounds (Li and Jain, eds, 2009). The F0 is closely related to pitch, which is defined as our perception of fundamental frequency. That is, the F0 describes the actual physical phenomenon, whereas pitch describes how our ears and brains interpret the signal, in terms of periodicity. Despite the difference, for simplicity, we use these two terms interchangeably in this paper.

speech (Murphy and Akande, 2005). Higher HNR means that the voice sounds less rough.[7] Both the shimmer and jitter are measures of voice perturbation. The shimmer measures the variation of voice amplitude and the jitter measures the variation of local pitch. Jitter is also known as frequency perturbation and refers to the minute involuntary variations in the timing variability between cycles of vibration. Jitter values in normal voices range from 0.2 to 1 percent. The formant is the broad spectral maximum that results from an acoustic resonance of the human vocal tract. It is a characteristic of the resonances of the space and a concentration of acoustic energy around a particular frequency in the speech wave. Each formant corresponds to a resonance in the vocal tract. The formant with the lowest frequency is called F1, the second F2, the third F3, and the fourth F4.

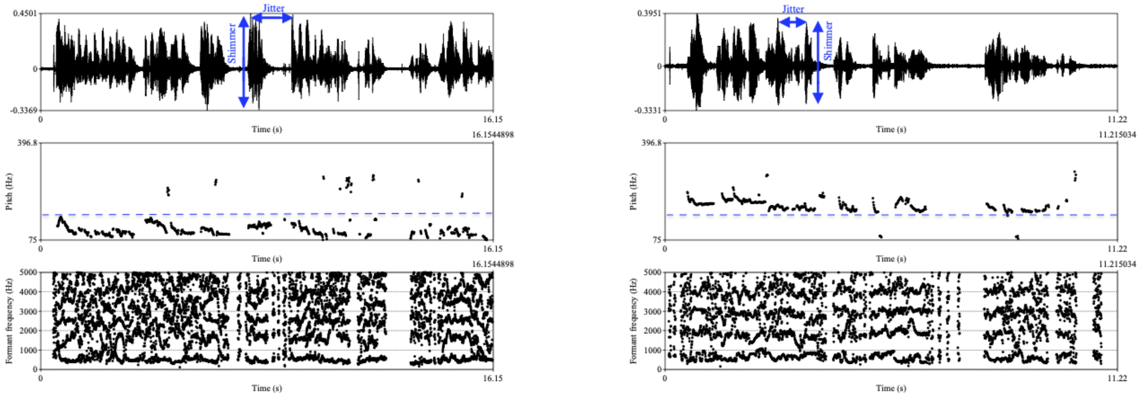
Figure 2 shows the graphic representation of our acoustic features extracted from two defendants' voice in RT, a male's voice on the left and a female's voice on the right. The top panels show voice wave with jitter (horizontal distance) and shimmer (vertical distance). The jitter and shimmer are calculated by averaging the absolute difference between a period and the two neighbors. The middle panel depicts pitch that ranges between 75 Hz and 400 Hz. While the average pitch of the male's voice is below the dashed line (155Hz), the average pitch of the female's voice is above it. In general, the average pitch of male's voice is lower than female's. Finally, formants displayed as dard bands can be seen very clearly in a wideband spectrogram. The lower panel shows the first, second, third and fourth formants. The darker a formant is reproduced in the spectrogram, the stronger it is.[8] A more detailed explanation of these features can be seen in Table B1 in the Appendix. Table B1 also includes the mean and standard deviation of the numerical values of the voice features for all subjects

[7] HNR is an assessment of the ratio between periodic components and non-periodic component comprising a segment of voiced speech. The periodic component arises from the vibration of the vocal cords and the non-periodic component follows from the glottal noise. The evaluation between the two components reflects the efficiency of speech, i.e., the greater the flow of air expelled from the lungs into energy of vibration of vocal cords, the higher the HNR is.

[8] It represents the more energy in the voice, or the voice is more audible.

in our two samples.

Figure 2: Voice Waveform with Acoustic Features



((a)) A Male's Voice

((b)) A Female's Voice

Additional Features

Along with the acoustic features, we also extracted the emotion from voice. We used Voice Emotion AI, Empath[9], which is the only professional emotion-extraction tool available free of charge. Empath is able to automatically detect four emotions: joy, calmness, anger, and sorrow from real-time speech in any language, even in high-noise environments. Joy and calmness are positive emotions whereas anger and sorrow are negative. We were also able to extract energy from voice using Empath. These four emotions and the energy from voice are highly correlated (see Table B5 for the correlation table). We selected two most distinct variables among the four: calm (positive, low intensity) and anger (negative, high intensity) for our analysis. These emotion parameters were extracted from the first 5 seconds of each audio clip.

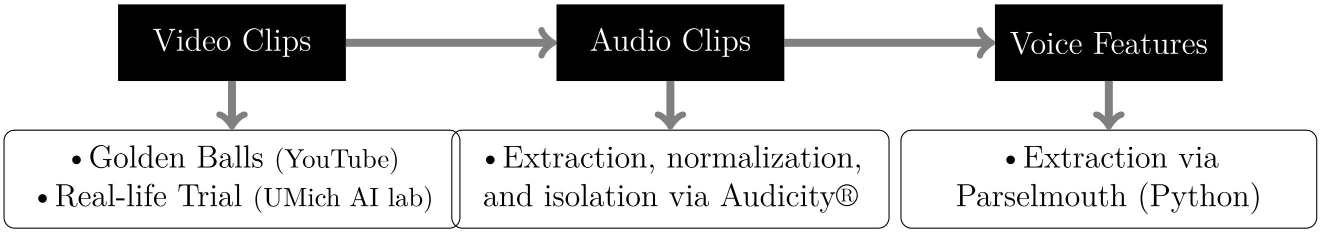
We used IBM Speech to Text algorithm to extract the initial transcripts of each audio and manually fixed the incorrect transcripts. Next, we performed Valence Aware Dictionary and

[9] See https://www.webempath.com/

Sentiment Reasoner (VADER) sentiment analysis to our transcripts. VADER is a lexicon and rule-based sentiment analysis tool that works well on texts. It takes in a string and returns a dictionary of scores in each of four categories: negative, neutral, positive, and compound. The positive, negative, and neutral scores represent the proportion of text that falls in these categories, and compound is computed by normalizing those scores between -1 and 1. Finally, we extracted additional features that could be associated with cooperative or deceptive behaviours in the GB game. These include whether the opponent was female, whether the player initiated the conversation, and the amount of money in the jackpot. In the text analysis, we also extracted the unigram from the bag of words analysis of each transcript and kept the top five words that were indicative of cooperation or defection, such as split, money, trust, promise, etc.

To summarize: (1) We collected the available video clips via YouTube for the GB and the video clips for RT from the University of Michigan AI laboratory; (2) We subsequently extracted and cleaned audio clips from the video clips using Audicity®; (3) We extracted: A. Voice features (i.e., pitch, intonation, etc.) via the Python Parselmouth library, B. Emotion from voice via Empath API, and C. Sentiment from text via Python VADER library. These steps are depicted in Figure 3.

Figure 3: Process Flow Diagram


Video Clips

Audio Clips

Voice Features

• Golden Balls (YouTube)
• Real-life Trial (UMich AI lab)

• Extraction, normalization, and isolation via Audicity®

• Extraction via Parselmouth (Python)

2.4 Results

2.4.1 Summary Statistics

Of the 236 audio clips in the GB data and 108 audio clips in RT, 127 and 65, respectively originated from females. In general, and consistent with our expectations, women spoke at a high pitch than men.[10] In GB the mean of females' pitch was 227.8 Hz while the males' was 211.13 Hz. The pitch in RT ranged from 118.01 Hz for males to 183.49 Hz for females. The higher pitches in GB compared to RT are likely due to the background music in the TV show, which had the effect of inflating the overall pitch of the audio recordings, this is known as the Lombard effect.[11] Intonations of both gender in GB were higher than those in RT because of the Lombard effect, the overall HNR (roughness of voice), shimmer (amplitude local variations), and jitter (local pitch variations) were comparable between GB and RT. Table 1 shows the summary statistics of acoustic features for both datasets separated by gender.

Table B2 in the Appendix shows the data reflecting the number of female and male contestants in GB and defendants in RT who we classified as cooperative or deceptive and their associated voice features. As previously mentioned, we labeled the statements of those in GB who chose to split as cooperative, and those who chose to steal as deceptive. Similarly, the depositions of those defendants in RT who later were shown to have told the truth were labeled truthful whereas those who lied were labeled as deceptive. Overall, in GB 58% of females and 53% of males were cooperative. In RT 32% of females and 74% of males were truthful.

[10] An adult woman's average range is from 160 Hz to 300 Hz, while a man's ranges from 60 Hz to 180 Hz (Re et al., 2012).

[11] The Lombard effect is the involuntary tendency of speakers to increase their vocal effort, such as increasing pitch of voice, when speaking in loud noise to enhance the audibility of their voice(Lane and Tranel, 1971).

Table 1: Summary Statistics

	Golden Balls (n=236)		**Real-life Trial (n=108)**	
Variable	Female	Male	Female	Male
Pitch	227.8	211.13	183.49	118.01
	(29.72)	(52.9)	(24.79)	(23.87)
Intonation	73.59	108.48	39.32	32.95
	(18.76)	(23.6)	(12.37)	(20.47)
HNR	8.65	6.9	9.26	7.75
	(2.19)	(2.0)	(2.25)	(2.43)
Shimmer	0.065	0.086	0.066	0.067
	(0.016)	(0.019)	(0.018)	(0.02)
Jitter	0.014	0.021	0.014	0.014
	(0.006)	(0.008)	(0.005)	(0.006)
N	127 (54%)	109 (46%)	65 (60%)	43 (40%)

2.4.2 Machine Learning Prediction Results

For our analysis, we employ two machine learning methods that are effective at classifying groups in small sample settings. These are decision tree and random forest. Decision tree creates a tree diagram that allows us to visualize features and interpret the results. However, decision tree can be unstable as small variations in the data might result in a large change in the tree diagram and it is also prone to overfitting. Random forest, a combination of multiple decision trees, can reduce the risk of overfitting, and produces more accurate results (please see Machine Learning Methods section in Appendix B for more details). Thus, we use decision tree for visualizing the model (see Decision Tree section in Appendix B) and random forest for predicting outcome variables. Due to the heavy imbalance sample in our datasets (especially between deception and truthful in RT), we use SMOTE (Synthetic Minority Oversampling Technique) method to resampling each of our datasets in which we try to simulate a balanced dataset by synthetically oversampling the minority group rest (see more details of SMOTE technique in Appendix B). This method is proven to be efficient in balancing the data by selecting similar records and altering that record one column at a time by a random amount within the difference to the neighboring records (Chawla et al., 2002). We test the validation accuracy of these machine learning models with the benchmark accuracy of naïve guess which represents a situation where we guess the occurrence of cooperation in GB or deception in RT by 50% of chance (similar to flipping a coin). Lastly, we iterate each of our models 1,000 times to find the range of the prediction accuracy.

Table 2 shows the results of our machine learning models using acoustic features, including pitch, intonation, HNR, shimmer, jitter, and formants, to predict cooperation in GB and deception in RT. The top panel presents the results for GB, left for female and right for male. The naive guess shows the initial probability of correctly predicting variable of interests, which is always around 50%. The accuracy shows the average prediction accuracy of our model with 1000 iterations. We use 75% training set and 25% test set for all the models. Utilizing random forest model with voice parameters from GB, we are able to predict cooperation with an average accuracy of 59.2% for females and 58.2% for males. This is an improvement in prediction accuracy of 8 - 9% compared to the benchmark of naïve guess. Using the same model with the RT dataset, we can predict deception with a higher average accuracy of 76.3% for females and 78.9% for males. As expected, our random forest models outperform our decision tree models in every instance.

These results are robust in 1,000 iterations and the t-test shows that all the machine learning models are significantly better than the naïve guess. Interestingly, adding text features such as text sentiment in RT or text sentiment and specific words (unigram) in GB doesn't improve the prediction accuracy in most of our models. Table B3 in the Appendix shows that text features did not add any extra information in predicting the cooperative behavior of females in GB and deceptive behaviors of both females and males in RT. There is even a 3% decrease in predicting deception of females in RT after adding the text sentiment. The only exception is predicting cooperation of males in GB. In this case, text features improve prediction by 7%.

In summary, our results validate the idea that voice contains some inherent characteristics that can help us identify cooperative and deceptive intent in real high-stakes interactions.

Table 2: Machine Learning Model: Voice Features

Model	Accuracy	SD	Model	Accuracy	SD
Golden Balls: Female			***Golden Balls: Male***		
Naive Guess	0.505	0.083	Naive Guess	0.495	0.093
Decision Tree	0.570	0.079	Decision Tree	0.555	0.092
Random Forest	0.592	0.074	Random Forest	0.582	0.086
Real-life Trial: Female			***Real-life Trial: Male***		
Naive Guess	0.504	0.101	Naive Guess	0.5	0.125
Decision Tree	0.726	0.102	Decision Tree	0.697	0.118
Random Forest	0.763	0.089	Random Forest	0.789	0.104

Note: These machine learning models to predict truthful statements use voice features only. The naive guess shows the initial probability of correctly predicting variable of interests. All procedures are repeated 1,000 times using random state sampling with Synthetic Minority Oversampling Technique (SMOTE). The accuracy is an average prediction accuracy after 1,000 iteration. We use 75% training set and 25% test set. Voice features include pitch, intonation, HNR, shimmer, jitter, and four formants.

2.4.3 Acoustic Features Identification

The ML models show that acoustic features of voice are effective in predicting cooperative and deceptive behaviors. We will now explore which of the acoustic features are significantly associated with cooperation or deception.

$$Y_i = \beta_0 + \beta_1 \, Acoustic\, Features_i + \varepsilon_i \tag{1}$$

Equation (1) shows our basic logistic regression model using acoustic features to predict the cooperation and deception. In this model, the dependent variable Y_i represents the binary value for each individual i with 1 representing cooperative behavior and 0 representing deceptive behavior in GB, while in RT 1 represents deceptive behavior and 0 represents truthful behavior. $Acoustic\, Features_i$ is a vector of variables representing voice parameters including pitch, intonation, HNR (harmonic-to-Noise Ratio), shimmer, jitter, and formants.

$$Y_i = \beta_0 + \beta_1 Acoustic\,Features_i + \beta_2 Emotion\,Features_i + \beta_3 TextSentiment_i + \beta_4 x_i + \varepsilon_i \quad (2)$$

Equation (2) presents our full model with various controls. $Acoustic\,Features_i$ is identical to Equation (1). $Emotion\,Features_i$ is a vector of variables representing emotions extracted from voice, including positive emotion (calm) and negative emotion (anger). $TextSentiment_i$ represents the compound factor of sentiment extracted from text transcripts that ranges from -1 to 1. The variable x_i is a vector of variables that are specific to GB, including unigram from text transcripts, the amount of jackpot, whether the contestant initiated the conversation or not, and the sex of opponent. Those factors are presumed to have an effect on the outcome variable. Last, in our logistic analysis, we also cluster the standard errors within each game in GB and each individual in RT as there might be some correlation within each game in GB and each individual in RT.

Table 3 shows the logistic regression results of GB. The first two columns show the results of basic voice model using acoustic features to predict cooperative behavior for females and males. And the last two columns show the results of the full model after adding text sentiment, emotions from voice, and other controls such as unigram and the sex of the opponent, etc. For simplicity, we excluded the results of shimmer, jitter, and formants in this table. The coefficient for males' pitch and roughness (i.e., HNR) show an association between acoustic features and cooperative behavior. The higher the pitch is, the more likely a male contestant is to cooperate. The result suggests that changing from the lowest register of Baritone (common male voice range) to the lowest register of Tenor (high male voice range) would increase the propensity of cooperation by roughly 10%. While the intonation does not play a role in predicting cooperation, the roughness of males' voices, which is higher in older men, does. More specifically, lower HNR (more roughness) is linked to more cooperation.

This result is in line with Van den Assem et al. (2012) who find that male contestants in GB become increasingly cooperative as their age increases. The last two columns show that these results remain after adding text sentiment, emotion of voice, and others controls. Column (4) in Table 3 shows that the sentiment from text is also associated with cooperative behavior. Negative sentiment extracted from text transcripts correlates with a higher likelihood of male's cooperative behavior. Interestingly, a male player is more likely to cooperate when facing a female opponent.

Table 3: Logit Results: Cooperative Speech in Golden Balls

Dep. Var: Cooperation	(1) Female	(2) Male	(3) Female	(4) Male
Pitch	-0.001	0.005**	-0.002	0.004**
	(0.002)	(0.002)	(0.002)	(0.002)
Intonation	-0.001	0.0002	-0.001	0.0005
	(0.003)	(0.002)	(0.003)	(0.002)
HNR	-0.005	-0.074*	0.003	-0.069**
	(0.027)	(0.037)	(0.026)	(0.034)
Text Sentiment			0.134	-0.291***
			(0.099)	(0.094)
Emotion: Calm			-0.000	0.004
			(0.005)	(0.003)
Emotion: Anger			0.007	0.015
			(0.009)	(0.01)
First Talker			-0.014	-0.115
			(0.1)	(0.073)
Female Opponent			-0.009	0.157**
			(0.082)	(0.072)
N	127	109	127	109

Notes: The column in each panel shows the average marginal effects with standard errors in parentheses; Standard errors are clustered by game. The full set of acoustic features used in all the models includes pitch, intonation, HNR, shimmer, jitter, and formants. The last two panels also control uni-gram from text transcript. Significance: * p<0.1, ** p<0.05, *** p<0.01.

Table 4 shows the logistic regression results of RT using acoustic features to predict deceptive behavior. Similarly, the first two columns present the results of basic voice models and the last two show the results of full models. In this data set, the pitch is positively associated with deception for both females and males. Higher pitch indicates deception. The result suggests that changing from the lowest register of Baritone (common male voice range) to the lowest register of Tenor (high male voice range) will increase the propensity of deception by 19%. Interestingly, we also find a positive correlation between pitch and cooperation for males in GB as mentioned above. This implies that a higher pitch of males' voice indicates both cooperation and deception in different contexts. The divergent results may stem from the different nature of the two contexts. During deposition in a trial, the defendant provides a narrative of what happened, while in the game show contestants interact back and forth to persuade the other to take an action. Past studies have shown that speakers adopt different styles of speaking in different contexts. For instance, conversational speech is distinct from read speech, with differences occurring in the speech rate, overall F0, and range of F0 (Hazan and Baker, 2010). Moreover, when facing an intense courtroom situation and in front of judges, people are mentally forced to suppress the emotion and tone down their voice. Conversely, in the game situation, people act more emotionally, such as increasing their pitch of voice. Last, the background music of the game show might also play a role in affecting acoustic features such as pitch during conversation. In the presence of loud competing noises, speakers will increase their voice effort, which has been linked to higher intensity and F0 as well as longer vowel duration (Summers et al., 1988; Traunmüller and Eriksson, 2000).

We also find that intonation is a good predictor of deception for both females and males in RT. Higher intonation is linked to a higher probability of being deceptive. While HNR (the roughness of voice) is not relevant in predicting deception in females' voices, it is associated with deception in males' voices, with a negative coefficient. Our results are also robust with the inclusion of emotions from voice and text sentiment.

Column (3) of Table 4 shows that text sentiment has a negative association with deception in females' voice. It implies that the more positive the text is, the less likely it is deceptive,

Table 4: Logit Results: Deceptive Speech in Real-life Trial

Dep. Var: Deception	(1) Female	(2) Male	(3) Female	(4) Male
Pitch	0.008**	0.009***	0.007**	0.01***
	(0.003)	(0.002)	(0.003)	(0.003)
Intonation	-0.014***	-0.009***	-0.015***	-0.009***
	(0.05)	(0.003)	(0.005)	(0.003)
HNR	-0.047	-0.159***	-0.095	-0.162***
	(0.056)	(0.048)	(0.065)	(0.053)
Text Sentiment			-0.193*	-0.049
			(0.116)	(0.154)
Emotion: Calm			-0.006*	0.006
			(0.003)	(0.01)
Emotion: Anger			0.062**	0.009
			(0.027)	(0.008)
N	65	43	65	43

Notes: The column in each panel shows the average marginal effects with standard errors in parentheses; Standard errors are clustered by individual (RT). The full set of acoustic features used in all the models includes pitch, intonation, HNR, shimmer, jitter, and formants. Significance: * p<0.1, ** p<0.05, *** p<0.01.

while such effect is not found in males. Moreover, calm and anger are also relevant predictors of deception for female in RT. Less calm or more anger is associated with deception. We do not find such relationship for male in this dataset.

2.4.4 Transferability

People adjust their languages according to the contexts. This makes it difficult to generalize the model of text features that is effective in predicting behaviors such as cooperation or deception in one context to another. On the other hand, considering that voice is unique for each person and primarily constrained by physical structures such as vocal tract, it is more difficult to modify voice in various manners. Thus, voice could in principle allow us to see whether features predictive of deceptive behaviors in one context are predictive in a different context.

Table 5: Cross-context Validation: Voice Features

Model	Accuracy	SD	Model	Accuracy	SD
GB to RT: Female			***GB to RT: Male***		
Naive Guess	0.499	0.055	Naive Guess	0.503	0.068
Decision Tree	0.54	0.07	Decision Tree	0.449	0.085
Random Forest	0.534	0.054	Random Forest	0.413	0.06
RT to GB: Female			***RT to GB: Male***		
Naive Guess	0.501	0.041	Naive Guess	0.499	0.048
Decision Tree	0.508	0.041	Decision Tree	0.501	0.042
Random Forest	0.513	0.035	Random Forest	0.505	0.033

Notes: The naive guess shows the initial probability of correctly predicting variable of interests. All procedures are repeated 1,000 times using random state sampling with Synthetic Minority Oversampling Technique (SMOTE). The accuracy is an average prediction accuracy after 1,000 iteration. We use 90% of the source and target set per iteration. Voice features include pitch, intonation, HNR, shimmer, jitter, and four formants.

With the advantage of having two different datasets available, we are able to test the transferability of voice in predicting deceptive behavior. In GB, the voice of contestants who chose "steal" was labeled as deceptive and in RT, the testimony of defendants or witnesses was labeled as deceptive according to the police investigations and the final verdict. We assume that the acoustic features that can predict deception in GB can be used to predict deception in RT, or vice versa. However, contradicting with our prior assumption, our results show that people voluntarily modulate their voice according to the situations. As section 2.4.3 shows that the cooperative voice in GB is associated with deception in RT. For instance, a higher pitch indicates cooperation in GB and deception in RT. A lower HNR of males' voice is an indicator of cooperation in GB and also an indicator of deception in RT. With those results, it is not surprising that when using acoustic features of males' voice in GB to predict deception in RT, our machine learning model perform poorly, and vice versa (see Table 5). Similarly, when using acoustic features of females' voice in GB to predict deception in RT, our model is only slightly better than naive guess (50%). Those results suggest that speakers change their voice according to the contexts. The acoustic features linked to deception in one context don't necessarily transfer to another context.

2.5 Conclusion and Discussion

Communication is more than cheap talk. Even informal talk can affect behaviors such as cooperation or deception in human strategic interactions. Researchers have analyzed text messages and scripted conversations to identify features that are indicative of a person's willingness to cooperate or to deceive. Despite the usefulness of studying verbal aspect of communication like text, much information in a message is contained in the non-verbal aspects of communication, such as voice, facial expressions, and posture. In this paper, we depart from the conventional approach of analyzing scripted messages or text. Instead, we focus on the voice. Unlike written text which can be modified to hide intentions, and scripted messages that are usually specific to a context, voice can capture subconscious processes that can reveal intent. In this paper, we use machine learning models to determine whether acoustic features of the voice such as pitch, intonation, and roughness can help us predict cooperative and deceptive behavior in real high-stakes interactions.

Like Turmunkh et al. (2019), we collected data from Golden Balls (GB), a British TV show that mimics a prisoner's dilemma with pre-play communication. We also collected data from Real-life Trial (RT), a database that includes court depositions of real defendants. Using random forest, we show that acoustic features predict deceptive behavior in RT with 76% and 79% average accuracy for females and males, respectively. This signifies a 26% to 29% improvement in prediction accuracy compared to the naïve guess. Surprisingly, adding text sentiment or emotion from voice does not add any extra information to the prediction model. In GB, our model reaches 59% and 58% average prediction accuracy for female and male contestants, respectively; that is a 8% to 9% improvement from the naive guess. The comparatively lower accuracy in GB is partly due to the fact that contestants' conversations in the TV show were recorded over background music.

Our analysis also identifies specific acoustic features that are associated with deception in RT and cooperation in GB. We find that higher pitch in both females and males is associated with higher likelihood of deception in RT. Interestingly, in GB male's higher pitched voices are also associated with more cooperation. The apparent contradiction may stem from the different nature of the two contexts. During deposition in a trial, the defendant provides a narrative of what happened, while in the game show contestants interact back and forth to persuade the other to take an action. In line with observations of past studies (Hazan and Baker, 2010), speakers adopt different styles of speaking in different contexts. For example, when facing a silent courtroom situation and in front of judges, people are mentally forced to suppress the emotion and tone down their voice. Conversely, in the TV game show contestants talk over music and competing noises. Under these conditions, speakers will increase their voice effort, which has been linked to higher intensity, F0, as well as longer vowel duration (Summers et al., 1988; Traunmüller and Eriksson, 2000). We also find that higher intonation is associated with lower likelihood of deception in RT. We do not find any relationship between intonation and cooperation in GB. Finally, the roughness of males' voice (i.e., HNR) is also informative for both cooperation in GB and deception in RT. Our results are robust after controlling for sentiment extracted from text, emotion from voice, and other variables such as the sex of the opponent in GB.

Given the above-mentioned results, it is not surprising that our model performed poorly when testing for feature transferability. Indeed, the random tree model shows that acoustic features of deception in RT do not predict deception in GB (and vice-versa). This finding suggests that acoustic features linked to deception in one context do not necessarily transfer to another context.

To summarize, this paper contributes to the existing knowledge in several ways. First, we show that acoustic features such as pitch predict deception and cooperation in real life interactions. Second, we show that voice features predictive of deceptive behaviors are context dependent. All in all, using data extracted from real life high-stakes settings, we were able to show that the acoustic features of the voice alone can predict cooperative intent in a high-stakes prisoner's dilemma game and deceptive communication in a high-stakes court deposition. This paper opens up new avenues for research on the role of non-verbal voice communication in revealing likely behaviors. Compared to facial and body recognition technologies, voice is less intrusive to collect, and simpler to analyze. Voice is as unique to an individual as a fingerprint and can reveal emotion, personality and as we have shown here, intent to cooperate and deceive. We hope that this research will motivate follow-up studies that utilize the voice as data.

Chapter 3

Trick for a Treat: The Effect of Costume, Identity, and Peers on Norm Violations[1]

Coauthored with Matthew Gomies, Narek Bejanyan, Zhou Fang, Jason Justo, Li-Hsin Lin, Rainita Narender and Joshua Tasoff

3.1 Introduction

Clothing serves the social function of communicating information about the wearer to others. The economic importance of wearing the right clothes has led to memorable proverbs such as "clothes make the man", "dress for success", and "dress for the job you want and not the job you have". While clothes clearly operate externally by communicating information to others, clothes may also have an internal effect by influencing one's own sense of identity. Militaries dress their soldiers in uniforms as part of their socialization (Wakin, 2000; Akerlof and Kranton, 2005). Adam and Galinsky (2012) find that wearing a lab coat increases performance on attention-related tasks and school uniforms have been found to reduce disciplinary referrals (Sanchez et al., 2012). While the mechanisms for these effects are not entirely clear, the results suggest that clothes can affect people's behaviors, perhaps through affecting one's sense of identity.

We test this hypothesis by recruiting trick-or-treaters — children in costumed garb — during the American holiday of Halloween. We consider this a boundary condition for the effect of clothing on identity and behavior, as it is the day of the year in which participants are dressed to the greatest extremes. Moreover, trick-or-treaters often use their costume to take

[1] The published version of this paper is available here: https://doi.org/10.1016/j.jebo.2020.09.004

on the identity of specific characters from film or television, and these specific assumed identities may have particularly salient effects on behavior. In addition, the function of costumes for festivals and holidays, on its own, merits scientific study. Consumers spent $9 billion in 2018 on Halloween products (National Retail Federation, 2018) and many cultures around the world have developed traditions in which costume-wear plays an important role. Interestingly, costume-wear often develops alongside traditions of norm-violations. Halloween evolved from the Celtic holiday of Samhain (Winkler and Winkler, 1970), in which costumes were used, in part, to hide one's identity during "tricks" or pranks (Miller et al., 1991). In Venice, the tradition of wearing masks during Carnival developed alongside activities that would otherwise be norm violations, such as mingling with other social classes, gambling, having clandestine affairs, reveling, and illicit activity (Walker, 1999; Burke, 2005).

Thus, it is historically and culturally apropos to measure the effect of Halloween costumes on ethical behavior. We use the lying game of Fischbacher and Föllmi-Heusi (2013) as our experimental paradigm. Trick-or-treaters privately roll a 6-sided die. If they report 1–5, they receive one candy, and if they report a 6, they receive the one candy and an additional bonus candy. If everyone tells the truth then the distribution of reported numbers would be uniform. Sixes reported in excess within a sample of individuals can be interpreted as evidence of lying for personal gain.

We manipulate three dimensions of the experimental conditions. First, we vary the stakes to price lying behavior. In the high-stakes condition, a reported six earns two bonus candies (three total) instead of the one bonus candy. Second, we vary the beneficiary of the lie to test whether lying for others is normative. On the one hand, incurring a psychic cost of lying to benefit someone else may be unappealing for some, but on the other hand, violating a norm in order to benefit someone else may be itself perceived as normative as there is no possible intimation of selfish behavior. In the baseline condition, the beneficiary is one's self. In the "other" condition, reporting a six earns someone in the next group of trick-or-treaters an additional candy. In the “both” condition, the trick-or-treater rolls two dice, the first affects one’s own payoff, and the second affects the payoff of a trick-or-treater in the next group. The both condition allows us to measure whether behavior in the “self” (“other”) condition spills over to behavior in the “other” (“self”) condition.

Third, we vary the degree of salience of one’s costume to observe how it affects the trick-or-treater’s ethical behavior. The variation in Halloween costumes leads to a natural separation in costumes between heroes and objects of admiration on the one hand, and villains and creatures of a wicked nature on the other. For example, the most popular costumes in our sample are (in order of popularity) a unicorn, Spiderman, Batman, Master Chief of the video game *Halo*, evil clown, vampire, Jason from *Friday the 13th.* The first four would be considered admirable by most people, while the latter three would be considered wicked by most people. In the treatment condition, we ask the trick-or-treater who they are dressed as, whether that character is a “good guy or bad guy”, and whether that person does “good things or bad things”. This is intended to draw salience to the person’s costume and the character’s ethical orientation. In the control condition, we ask the same questions but after the participant has already reported their dice rolls.

Additionally, natural variation in participant age and arrival affords us the ability to answer two other thematic questions. (1) Does age affect lying? This is an important question, as a vast developmental psychology literature shows that children’s cognitive abilities and behaviors mature at specific ages. (2) To what extent is lying in children influenced by peers? The experiment was run at a house on a street that hosts large crowds of trick-or-treaters. Participants lined up and were allowed to advance to the porch in groups of 10, where upon they were told the rules of the game. While participants advanced to the door of the house one-by-one, participants behind them could likely hear their reports. Because we recorded the order in which participants lined up, we can measure the effect of reporting a six on subsequent participants’ reports. Specifically, we test the extent to which trick-or-treaters are affected by same gender versus opposite gender peers. These two questions fall within the paper's broader theme of identity, in this case relating specifically to the trick-or-treater's age and gender identities.

We found frequent occurrence of six with 40% of participants reporting a six in the baseline condition, which is approximately 23 percentage points more sixes than by chance alone. Stakes had no effect on the number of reported sixes. The other-condition strongly reduced the frequency of reporting a six. Many trick-or-treaters did not view the cost of lying to be worth helping out an anonymous stranger. In the both-condition, the occurrence of reporting a six to benefit one's self decreased, while the occurrence of reporting a six to benefit other's did not change. The result suggests that participants' honesty in reporting for others spilled over to reporting for self but not vice versa.

Next, we test the effect of costume salience on participant's behavior. Costume choice is endogenous: the correlation between costume and lying would not necessarily indicate a causal relationship. However, increasing the salience of a costume is random and it is hypothesized to have different effects for those who are dressed as characters of admiration (heroes or creatures of beauty), and those who are dressed as wicked characters. We hypothesized that drawing attention to the ethical orientation of one's assumed identity would lead to behavior consistent with that identity. That is "good guys" whose costumes are rendered salient would lie less than "good guys" whose costumes are not made salient and "bad guys" whose costumes are rendered salient would lie more than "bad guys" whose costumes are not made salient. In fact, we found the opposite. The salience condition caused good guys to lie more, by about 12 percentage points, than good guys in the control condition. Bad guys with the salience of their costumes lied significantly less, by about 27 percentage points less, than bad guys in the control condition. We offer two possible interpretations of these results. The results are consistent with a moral licensing effect for good guys (Secilmis, 2018; Lasarov and Hoffmann, 2018). Making "good guy" salient may have made individuals feel more justified in committing a norm violation, and the reverse effect may have operated on "bad guys". Alternatively, the effect may be driven by a feeling of being monitored. The trick-or-treaters declared to an observing adult whether they were a "good guy" or a "bad guy". This may have changed the trick-or-treaters' perceptions about the extent to which they were being monitored. Perhaps self-declared "good guys" felt as though they were putting themselves in our good graces, while self-declaring oneself as a "bad guy" felt like it would warrant greater monitoring from us. We cannot test between these two hypotheses but the physical structure of our experiment casts some doubt on the plausibility of the latter hypothesis.

Finally, examining the natural variation of our sample, we find an inverted-U pattern for the effect of age, with lying peaking at age 12. We find large and statistically significant peer effects. One additional person reporting a six out of a group of five participants increases the probability of reporting a six by about 5 percentage points. Interestingly, as we decompose this effect by gender, we find that most of the effect operates within gender. That is girls follow girls and boys follow boys. It appears that gender identity prominently moderates the peer effect.

We view our main contribution as showing that salience of one's clothed identity affects ethical behavior in the direction of a moral licensing/self-conscious effect. Additionally we find that gender identity influences behavior, with girls emulating girls, and boys emulating boys in their reporting behavior. We also provide evidence on whether lying for others is viewed as normative for children, and provide evidence for the existence of within-person spillover effects and age effects. In the next section we relate our study to the literature on lying, especially as a function of age, stakes, beneficiary and peers. We also discuss the literature on clothing, costumed festivals, identity and norm violations. The experiment design is described in Section **??**. Section 3.4 contains the results. Section 3.5 and 3.6 contain the Discussion and Conclusion, respectively.

3.2 Literature Review

Fischbacher and Föllmi-Heusi (2013) developed an ingenious experimental paradigm for the measurement of aggregate lying at the population level. In the original study participants roll a die privately and obtain a reward based on the result they report to the experimenter. They found that about 20% of participants lie to the fullest extent possible while 39% of them are fully honest. The experiment has generated numerous variants (see Abeler, Nosenzo and Raymond, 2019, for a meta-analysis). Notably, Cadsby, Du and Song (2016) run a variant in which they find that people lie to increase the payoff of an in-group member even though such a lie does not affect their own monetary payoff.

A few recent economic papers have examined lying in children. Bucciol and Piovesan (2011b) conduct a similar experiment in a childrens' summer camp with ages 5-15. They find no association between lying and age, however they only report a linear specification and their sample is small in their baseline condition (N=81). Glätzle-Rützler and Lergetporer (2015) find that 10 and 11 year olds are more likely to lie than 15 and 16 year olds, consistent with our age pattern. Maggian and Villeval (2016) run a somewhat different game in which lying can yield higher payoffs, with 7–14 year olds. They find that 9-10 year olds lie more than 7-8 year olds or 11-14 year olds. This is similar to our finding as we also have an inverted-U as a function of age, though our peak age is 12. Brocas and Carrillo (2019) find that middle-schoolers aged 11-14 lie significantly more than the other age groups (5-8 year olds, 8-11 year olds, 14-17 year olds, and undergrads), which confirms our results. They additionally provide evidence on lying as a function of age when the benefactor is another child. They find that only middle-schoolers exhibit lying in the aggregate, and they lie to reduce the other child's payoff! In aggregate, we find that our subjects lie to benefit others; the difference between these results may be due to design differences and differences in payoffs (for example, in their experiment, payment is strictly increasing in the die roll). Finally, on the topic of stakes Fischbacher and Föllmi-Heusi (2013) and Mazar, Amir and

Ariely (2008) do not find an effect of stakes on lying.

The research on peer-effects of norm-violations consistently finds that violations are contagious. Gino, Ayal and Ariely (2009) conduct a cheating experiment in which a confederate publicly announces that he has completed a task in an impossibly short amount of time. If the confederate is an in-group member, cheating goes up relative to the baseline. Interestingly, if the confederate is an out-group member (dressed in the clothes of a rival school) cheating goes down relative to the baseline. Diekmann, Przepiorka and Rauhut (2015) and Bicchieri, Dimant and Sonderegger (2020a) include conditions in lying experiments in which subjects are informed about the lying behavior of others. They find that information about the extent of lying in the population increases lying compared to the control condition. In a charitable contribution game, Bicchieri et al. (2020b) find similarly that learning about the empirical distribution of other's contributions causes less compliance with giving. Our results strengthen this literature, showing that direct observation of peers reporting a six increases reporting a six. Our results extend this literature by showing that contagion occurs within gender rather than across gender, in a sample of children. Consistent with the findings of Gino, Ayal and Ariely (2009), we show that the identity of the observed cheater moderates the contagion effect. They show it occurs via school affiliation, we show that the results extend to gender identity.

Abeler, Nosenzo and Raymond (2019) summarize the theory literature on the Fischbacher and Föllmi-Heusi (2013) paradigm. They categorize models into those that have lying costs only, those that have lying costs with a preference for conformity, and those that have lying costs and a preference for honest reputation. They conduct a Herculean meta-analysis of the experimental literature as well as an additional large experiment of their own and they find that only the last category of models, lying costs and a preference for an honest reputation, explains the pattern of results. Our results are consistent with this finding in the sense that (1) lying in our study is not maximal and this can be explained by lying costs. (2) Our salience effect — salience causes good guys to lie more and bad guys to lie less — could be recast as a preference for an honest reputation. The salience may make good guys feel like their ex-ante reputational capital is higher, and the opposite may be true for bad guys. If there are diminishing returns to reputation, then those who feel they have higher reputation are more likely to spend it for a sweeter physical reward. This could generate our result. Of course, we do not observe or directly manipulate self-perceived reputation in our study, so this latter point is speculative. We did not find economic models that relate to our other research questions.

To the best of our knowledge, no study has examined the effect of the salience of clothing on lying behavior. Research on the economics of clothing is limited but it has a long history going at least as far back as Veblen (1899). Veblen observed that restrictive or delicate clothing can be a costly signal indicating that the wearer is of the well-to-do "leisure class", since a laborer would be unable to function in such clothes. Clothing also communicates people's roles. For example, a concerned citizen can identify a police officer during an emergency thanks to the officer's uniform. Given that clothes communicate a person's identity to others, researchers have posited that clothes may also affect one's self-identity. Adam and Galinsky (2012) find that wearing a lab coat increases performance on attention-related tasks. School uniforms can reduce discipline referrals by 9.7% (Sanchez et al., 2012). Civile and Obhi (2017) randomly assign participants to wear police-style uniforms and then have them engage in an attention-related task. They find that uniforms cause participants to attend more to images associated with lower socio-economic status. The results establish that wearing clothes associated with particular roles in society, or identities, can causally affect behavior.

Wearing costumes on Halloween that assume the identity of specific characters has become a mainstream practice in American culture. A lab coat or a police officer's uniform have metonymic relationships with the institutions they represent. In contrast, many Halloween costumes reflect specific characters (e.g. Batman or Moana) as opposed to broader roles in society.[2] Thus it is not clear whether the greater specificity common in Halloween costumes can enable the kind of identification and change in behavior exhibited in the aforementioned studies. However, casual observation suggests that at least some level of identification occurs. Children often select characters that are aspirational or characters that they like to act out in pretend play. Pretend play often accompanies the wearing of costumes: Reys and Lukes swing their lightsabers against imaginary stormtroopers, Moanas steer their imaginary ships to safety, and vampires suck the imaginary blood of their real siblings (much to their parents' dismay). Indeed, it seems that an important reason why Halloween is so attractive to children is that it facilitates enjoyable roleplay.

Halloween evolved from the Celtic holiday of Samhain during which it was believed that spirits and souls of the dead would return to earth (Sterba, 1948). The festivities involved people going door-to-door in costume reciting verses in exchange for food (Linton and Linton, 1950; Ward, 1981). Later in the 16th century Scotland, revelers would wear masks or painted faces threatening to do mischief if they were not given food (Linton and Linton, 1950; Belk, 1990). Costume wearing and norm violations emerged contemporaneously in this context, and the co-emergence of the two institutions have arisen in other cultures as well. The wearing of costumes in Venice for Carnival originated alongside traditions of mischief-making and intermingling of social classes that were otherwise discouraged from mixing (Feil, 1998). We note several social functions that these costumes may have served. First, disguising one's outward identity while begging may have been a way for those individuals to avoid harm to their reputation and the associated shame. Second, disguising one's outward identity while violating any rule, be it legal or an implicit cultural norm, has the obvious benefit of avoiding punishment. Thus, disguises are complements with norm violations, the presence of one in a tradition increases the marginal value of including the other.

[2] We thank an anonymous referee for this important observation.

Masking one's identity may make one more prone to norm violation. In addition, research has shown that rendering specific aspects of one's identity salient can change one's propensity to violate norms. Recent studies find that bank employees become dishonest and clergymen become more honest when their professional identity is rendered salient (Celse and Chang, 2017). Cohn, Maréchal and Noll (2015a) show that increasing the salience of prisoner's criminal identity increases dishonest behavior. In these studies, making an aspect of the individual's identity salient promotes ethical behavior congruent with that aspect. However, a separate thread of research has established a potential countervailing force.

Moral licensing is the psychological phenomena in which boosts to self-image increase engagement in unethical behavior (Nisan and Horenczyk, 1990). Sachdeva, Iliev and Medin (2009) had participants write a short story about themselves or someone they knew using morally positive trait words (e.g., fair, kind) or morally negative trait words (e.g., selfish, mean). Participants assigned to write about themselves using positive traits donated the least out of the four conditions and those who wrote about themselves using negative traits donated the most. Khan and Dhar (2006) obtained similar results, finding that participants asked to imagine helping others donated less to charity than control subjects, and Mazar and Zhong (2010) show that people act less altruistically and are more likely to cheat and steal after purchasing green products than after purchasing conventional products. Jordan, Mullen and Murnighan (2011) had participants recall one's own moral or immoral past actions or another's moral or immoral past actions. They find that people who recalled their own immoral behavior reported greater participation in moral activities, reported stronger prosocial intentions, and showed less cheating than people who recalled their own moral behavior. Similarly, Clot, Grolleau and Ibanez (2014) find that participants who recalled their own moral actions subsequently cheated more to get a higher payoff than participants who did not recall their moral actions. We extend the literature by measuring whether drawing salience to one's costume affects the propensity to lie, differentially for those dressed as “good guys” vs. “bad guys”.

3.3 Research Design

3.3.1 Sample and Setting

Our experiment was conducted on the dark and stormy night of Halloween, October 31st, 2018, at the house of one of the authors in a suburb of Los Angeles. The neighborhood is a destination for trick-or-treating amongst nearby communities. A typical home in this neighborhood is visited by about 1,000 trick-or-treaters. Trick-or-treaters who approached the house between 6:00 and 9:30pm participated in the experiment. They were told that they could play a game to win candies. In total, 544 trick-or-treaters participated. Other participants of a spiritual and malevolent nature, may have participated, undimensioned and unseen, without our knowledge.

3.3.2 Experimental Procedure

The experiment proceeded as follows. An experimenter advertised to passing trick-or-treaters that they could play a game to win candies. Trick-or-treaters queued in front of the porch and were randomly assigned to two lines. One line led to the no-salience condition and the other line led to the costume salience condition. All subjects were given an ID card with a number (1-10) and were instructed that they would need their ID card to exchange for candy. All subjects were asked their age. A quietly observing experimenter recorded subjects’ answers discretely along with other information such as gender, whether parents accompanied them or not, time of the day, and the specific ID number. Those in the salience condition were asked additional questions, “Who are you today?”, “Is (answer to the previous question) a good guy or a bad guy?” and “Does (answer to the first question) do good things or evil things?” The subjects in the no-salience condition were asked the same questions but only at the very end of the experiment. Subjects advanced approximately in the order they arrived.

We use the lying game introduced by Fischbacher and Föllmi-Heusi (2013). Ten subjects were allowed to the porch at a time at which point we explained the rules of the game: they would role a 6-sided die in a paper cup. If they rolled a 1-5, they would get one candy; if they rolled a 6, they would win bonus candy. We also stated very clearly, "You don't need to show us the dice. Just tell us the number." Experimenters at the end of the line asked for the number and gave the promised number of candies.

3.3.3 Experimental Treatments

The experiment has 2×3×2 conditions: we varied the stakes (high vs. low), we varied who the beneficiary of the bonus candy was (self vs. other vs. both), and we varied the salience of subjects' Halloween costumes (no-salience vs. salience). As mentioned above, the salience condition was randomized at the individual level as trick-or-treaters approached the house. The stakes and beneficiaries were randomized by group. We cycled through the six conditions, alternating after every group of 10 subjects. We had a total of 55 groups.

We varied the stakes in order to price the effects of the other treatments. In both conditions, subjects receive one candy for reporting any number. In the low-stakes condition, they earned one bonus candy for reporting a six, and in the high stakes condition they earned two bonus candies for reporting a six. Though the stakes are low, a pilot study conducted the year prior suggested they would be adequate. In the pilot, the low-stakes condition paid 1 candy for a reported one through four, 2 candies for a reported five, and 3 candies for a reported six. We found an excess of fives and an even greater excess of sixes. In the high-stakes condition the rewards were 1 candy, 3 candies, and 5 candies. In that study we observed a significant effect of high stakes. We expected that the 1 additional bonus candy in the high-stakes condition would be sufficient.

In the “self” condition, the subject was the recipient of the bonus candy, and in the “other” condition the recipient of the bonus candy was the subject in the next group with the same ID number. In the “both” condition, subjects rolled one die for themselves and one die for a subject in the next group. As mentioned above, subjects in the salience condition were asked questions prior to the instructions, while subjects in no-salience condition were asked the same questions after the number was reported.

3.3.4 Additional Covariates

In our regressions we include additional covariates. We wish to test whether there is an age pattern in the propensity to lie. We censor ages from 4 to 19, to deal with outliers (there are the rare 50 year-old trick or treaters). We estimate a quadratic specification for age, allowing for curvature and a slope sign change.

Subjects were admitted to the porch in groups of 10 at which point they were handed dice and explained the instructions. The children were in close proximity to each other as they lined up to report their number. It would have been easy for children in line to overhear the reports and earnings of children in front of them. This accidental feature of the design allows us to identify peer effects. The exogenous arrival rate of trick-or-treaters creates variation in the composition of each group and our assignment of ID number creates variation in the sequential order of the reporting. We estimate a coefficient for a “peer proportion” variable defined as the proportion of previous participants in the group that reported a six. If the person is first in their group we define peer proportion as zero. We decompose the peer-proportion effect by gender. We define “female peer proportion” as the proportion of previous female participants in the group that reported a six, and define the “male peer proportion” variable in the analogous way for males. If there were no previous females in the group then “female peer proportion” is defined as 0 and if there were no previous males in the group then “male peer proportion” is defined as zero.

We include several covariates as controls to increase statistical power. In all statistical models in which age is not the primary focus the quadratic age variables are not included; we instead use a more flexible specification using 11 age categories. Ages 4 and less are the first category, each age from 5 to 11 are their own categories, and ages 12-14, 15-18, and 19+ are the last three categories. Some children are accompanied by their parents. Obviously, the presence of one's parents could have an effect on a child's propensity to lie so we include a parent indicator as a control. We also include a gender indicator as a control variable in all models.

3.3.5 Multiple-Hypotesis Testing

A potential weakness of our research design is the large number of variables of interest. Naïve multiple-hypothesis testing can inflate the chance of having statistically significant results by chance, even under the null hypothesis. To control for false positives we include a false discovery rate (FDR) analysis. We use the method of Benjamini, Krieger and Yekutieli (2006), which is the sharpened two-step FDR procedure. This procedure produces less false negatives than the original FDR procedure. Any coefficient displayed in a table is included in the set of hypothesis tests we use for our paper-wide FDR correction. The q-values are equivalent to p-values that have been adjusted for multiple hypothesis testing.

3.4 Results

3.4.1 Summary Statistics

Table 1 shows the summary statistics of our sample. Among a total of 544 subjects who participated in our experiment, about 53% are female, and 41% are accompanied by parents. The ages of our subjects range from 0 (baby) to 50 years old (as some parents also joined the game) with an average age of 9.46 years old. The very young children were accompanied by their parents. In the analysis, we bottom-code participants age at 4 years-old and top-code participant age at 19 years-old due to the small samples outside of this range. About 24% wore a self-reported bad-guy costume. The most popular costumes in order of popularity were: unicorn, Spiderman, Batman, the *Halo* video game main character, clown, vampire, and Jason from *Nightmare on Elm Street*.

Table 1: Summary Statistics

	mean	sd	min	max
Six	0.41	0.49	0	1
Six_Self	0.47	0.50	0	1
Six_Others	0.33	0.47	0	1
Group	27.99	15.72	1	55
Age	9.46	5.38	0	50
Parents	0.41	0.49	0	1
Self Condition	4.61	1.66	1	6
Others Condition	4.28	1.66	1	6
Female	0.53	0.50	0	1
High Stakes	0.51	0.50	0	1
Salience	0.50	0.50	0	1
Bad Guy	0.24	0.43	0	1
Good and Bad	0.02	0.13	0	1
Age Category	9.33	4.11	4	19
Six within Group	2.24	2.25	0	12
Six within Group_Female	1.14	1.46	0	10
Six within Group_Male	1.10	1.34	0	8
Observations	544			

The frequency of reporting a six within our sample is 41%, which is significantly above the probability of rolling a six at 16.7%. This implies that a substantial number of our subjects lied in the experiment. While traditional economic theory predicts that when there is no cost to lying, people would always report a six, our results show that at least some subjects are honest. The results are consistent with Fischbacher and Föllmi-Heusi (2013), with a substantial degree of lying.

Table 2: Raw Probabilities

	Outcome				
	Not Six		**Six**		**Total**
	No.	%	No.	%	No.
Stakes					
Low stakes	202	60	137	40	339
High stakes	211	60	142	40	353
Total	413	60	279	40	692
Beneficiaries					
Self	84	46	100	54	184
Other	104	61	67	39	171
Both	225	67	112	33	337
Both - Self	103	61	66	39	169
Both - Other	122	73	46	27	168
Total	413	60	279	40	692
Salience					
No Salience	203	59	140	41	343
Salience	210	60	139	40	349
Total	413	60	279	40	692
Gender					
Male	191	59	133	41	324
Female	220	61	141	39	361
Total	411	60	274	40	685
Costume					
Good Guy	279	61	181	39	460
Bad Guy	83	57	62	43	145
Total	362	60	243	40	605

Table 2 displays the raw totals and percentages of reporting a six by condition. The high stakes and low stakes conditions both have approximately 40% reporting a six. The differences between the beneficiary conditions is sizable. While 54% report a six in the self condition, only 39% report a six in the other condition. Reporting a six further drops in the both condition with 39% reporting a six for themselves and only 27% reporting a six for others. Reporting a six is just as likely in the no-salience and salience conditions with only a 1 percentage point separating the two. If there is a gender effect, it appears small, with boys reporting a six only 2 percentage points more than girls. True to their names, there does appear to be a difference between good guys and bad guys with six reports at 39% vs. 43%. To better assess the effects of the treatments we turn to regression analysis.

3.4.2 Main Results

We display the analysis of our treatments in Table 3. All columns report average marginal effects of a logistic regression. Each cell of the table contains four statistics. The number in the upper-left corner is the average marginal effect of the variable, the number in the lower-left corner in parentheses is the standard error. The number in the upper-right corner in italics is the naïve p-value of the average marginal effect and the number in the lower-right corner is the sharpened FDR q-value in square brackets. The q-value can be interpreted as a p-value that corrects for multiple hypotheses. Each cell of a table is considered a hypothesis test for the purposes of the paper-wide FDR corrected q-values.

Column (1) regresses an indicator for reporting a six as the outcome on the stakes condition, beneficiary conditions, gender, and the presence of parents. Additional unreported controls include age categories. The effect of the high-stakes condition is not significant. The "other" condition significantly reduces the frequency of reporting a six relative to the baseline "self" condition. Subjects are less willing to lie to benefit an anonymous stranger than they are if the benefit accrues to themselves. The negative coefficient in the "both"

Table 3: Reporting a Six – Logistic Regression with Average Marginal Effects

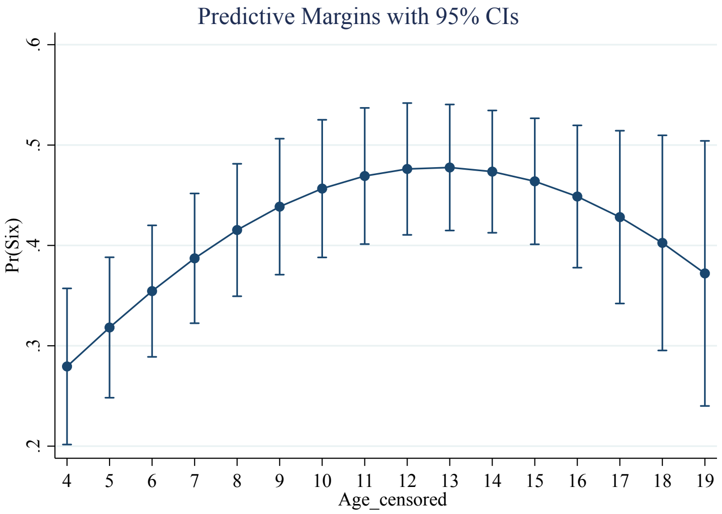
	(1)		(2)		(3)		(4)	
High Stakes	-0.005	*0.926*	-0.012	*0.809*	-0.002	*0.958*	-0.001	*0.981*
	(0.051)	[0.657]	(0.05)	[0.657]	(0.04)	[0.657]	(0.04)	[0.657]
Other	-0.131**	*0.025*	-0.137**	*0.018*	-0.115**	*0.015*	-0.117**	*0.013*
	(0.058)	[0.087]	(0.058)	[0.078]	(0.047)	[0.074]	(0.047)	[0.074]
Both	-0.13*	*0.062*	-0.14**	*0.046*	-0.092	*0.109*	-0.094	*0.107*
	(0.07)	[0.111]	(0.07)	[0.093]	(0.058)	[0.125]	(0.058)	[0.125]
Other x Both	0.012	*0.885*	0.018	*0.828*	-0.016	*0.832*	-0.017	*0.827*
	(0.082)	[0.657]	(0.082)	[0.657]	(0.076)	[0.657]	(0.076)	[0.657]
Female							0.008	*0.885*
							(0.052)	[0.657]
Age			0.067***	*0.002*				
			(0.021)	[0.039]				
Age^2			-0.003***	*0.009*				
			(0.001)	[0.065]				
Peer Proportion					0.268***	*0.0*		
					(0.059)	[0.001]		
Female Peer Proportion							0.277	*0.102*
							(0.17)	[0.125]
Female x Female Peer Proportion							0.036	*0.865*
							(0.213)	[0.657]
Male Peer Proportion							0.285***	*0.004*
							(0.1)	[0.046]
Female x Male Peer Proportion							-0.104	*0.461*
							(0.142)	[0.383]
Female Peer Proportion on Female							0.314***	*0.007*
							(0.117)	[0.06]
Male Peer Proportion on Female							0.18*	*0.08*
							(0.103)	[0.125]
N	685		685		685		685	
Clusters	55		55		55		55	

Notes: The dependent variable is whether the subject reported a six or not. All models are logistic regression with age categories, gender and parents included as controls. The first column in each cell reports average marginal effects with standard errors in parentheses. The second column in each cell reports naive p-values in italics and FDR adjusted q-values in brackets. Standard errors are clustered by group. Each cluster is a set of 10 subjects who were given instructions at the same time. Subjects in the "Both" treatment reported two outcomes. To analyze peer effects in Column (3) and Column (4), we exclude the first subject from each group whose probability of reporting a six should not be affected by anyone else. *** p<0.01, ** p<0.05, * p<0.1.

condition indicates that having two die rolls reduced the occurrence of reporting a six to benefit one's self. The results show no effect of gender.

In Column (2), we no longer control for age using categorical variables but instead include a second-order polynomial in order to fit an age trend. The coefficient on age is positive and significant while the coefficient on age^2 is negative, indicating that lying exhibits an inverted U pattern.[3] Figure 1 displays the predicted six-reports as a function of age with 95% confidence intervals. Note that the probability of reporting a six is significantly higher than 20% across the whole domain, which is above 16.7%, the expected frequency of reporting a six under truth-telling. These results indicate that lying happens at all age levels. The probability of reporting a six peaks at age 12 and subsequently decreases thereafter.

Figure 1: Reporting a Six as a Function of Age



Notes: Probability of reporting a six across ages. Error bars indicate 95% confidence intervals. Ages are censored at 4 and 19.

In Column (3) we test for peer effects on reporting a six. Recall, the "Peer Proportion" variable as the proportion of previous reports within the group that were six. The coefficient is positive and statistically significant with a p-value < 0.001. This provides evidence that

[3] Simonsohn (2018) suggests using the "two-line test" for testing whether two fitted lines have opposite signed slopes. The test confirms our inverted U finding. The results are available in Appendix C.

the probability of reporting a six is influenced by previous subjects in the group reporting a six. If one additional child out of five previous children reports a six, it leads to a 5.36 percentage point increase in the probability of a subsequent child reporting a six.

In Column (4) we decompose the peer effect by gender. “Female Peer Proportion” measures the proportion of female subjects before the subject in question who reported a six, and “Male Peer Proportion” does likewise for male subjects. While gender does not have a direct effect on one’s lying behavior, our data show a difference in how boys and girls are influenced by the other subjects in their group. The coefficient on “Female Peer Proportion” is the effect of previous female subjects on male subjects and it is positive but not significant. The coefficient of “Male Peer Proportion” is the effect of previous male subjects on male subjects, which is large, positive, and statistically significant ($p = 0.004$). As we turn to girls, the coefficient on “Female Peer Proportion” plus the coefficient on “Female × Female Peer Proportion”, 0.314 is positive, and significant ($p = 0.007$). The effect of boys on girls is “Male Peer Propotion” plus “Female × Male Peer Proportion” which is 0.18, and marginally significant at p=0.08. To summarize, boys emulate boys and girls emulate girls. We interpret this as evidence that children take their social cues within gender, tending to emulate others who share their gender identity.

In Table 4 we turn to the effects of the salience of subjects’ costumed identities. Column (1) contains the same regressions as from Table 3 except we add the additional indicator variables for self-reported “bad guy”, salience condition, and the interaction between the two. In this table we drop subjects who self reported as “both a good guy and bad guy” or as “neither a good guy or bad guy”. We hypothesized that subjects would behave more congruously with their costumed character in the costume salience condition. Because “good guys” and “bad guys” are expected to respond in opposite directions, the interaction term is essential. Instead, we find evidence for the opposite effect. The coefficient on salience is the effect on “good guys” and it is positive but insignificant. The coefficient on the

Table 4: The Effect of Priming on Reporting a Six – Logistic Regression with Average Marginal Effects

	(1)		(2)	
High Stakes	-0.044	*0.358*	-0.046	*0.333*
	(0.048)	[0.299]	(0.047)	[0.285]
Other	-0.122**	*0.044*	0.001	*0.99*
	(0.06)	[0.093]	(0.084)	[0.657]
Both	-0.138**	*0.048*	-0.14**	*0.044*
	(0.07)	[0.093]	(0.07)	[0.093]
Other x Both	0.001	*0.99*	0.006	*0.943*
	(0.09)	[0.657]	(0.088)	[0.657]
Salience	0.036	*0.435*	0.115**	*0.049*
	(0.046)	[0.37]	(0.058)	[0.093]
Bad Guy	0.089	*0.24*	0.214**	*0.019*
	(0.076)	[0.239]	(0.092)	[0.078]
Salience x Bad Guy	-0.169**	*0.036*	-0.268***	*0.003*
	(0.08)	[0.093]	(0.089)	[0.039]
Salience x Other			-0.168**	*0.044*
			(0.083)	[0.093]
Bad Guy x Other			-0.239**	*0.03*
			(0.11)	[0.092]
Salience x Bad Guy x Other			0.304	*0.112*
			(0.191)	[0.125]
N	600		600	
Clusters	55		55	

Notes: The dependent variable is whether the subject reported a six or not. All models are logistic regression with age categories, gender and parents included as controls. The first column in each cell reports average marginal effects with standard errors in parentheses. The second column in each cell reports naive p-values in italics and FDR adjusted q-values in brackets. Standard errors are clustered by group. Each cluster is a set of 10 subjects who were given instructions at the same time. Subjects in the "Both" treatment reported two outcomes. We managed to record the costume information for only 464 subjects. *** p<0.01, ** p<0.05, * p<0.1.

interaction term "Salience × Bad Guy" is the differential effect of costume salience on "bad guys" relative to "good guys" it is negative and significant at $p = 0.036$. The results are tantalizing yet unconvincing. We hypothesized that the relatively lower rates of lying in the "other" conditions may be attenuating the effect.

In Column (2), we decompose the effect of costume salience into "self" and "other" conditions. We include interactions between salience and "other", "bad guy" and "other", and the triple interaction between salience, "bad guy", and "other". The coefficient on "bad guy", which is positive and significant at $p = 0.019$, implies that those wearing "bad guy" costumes are about 21 percentage points more likely to lie to benefit themselves than those wearing a "good guy" costume. Costumes are self-selected so this could be either a selection effect, a causal effect of wearing the costume, or some combination thereof. The manipulation is not the costume but the the salience of the costume. The coefficient on salience is interpreted as the effect of the costume salience condition on "good guys" when the beneficiary is one's self. The coefficient increased relative to Column (1), and is now significant at $p = 0.049$. Interestingly, the coefficient on "Salience × Bad Guy", which is the differential effect of salience on "bad guys" relative to "good guys" when one's self is the beneficiary, is now larger in magnitude, negative and significant at $p = 0.003$. The overall effect of costume salience on "bad guys" when one's self is the beneficiary, is the sum of the coefficients on salience and "Salience × Bad Guy", which is -0.153 negative and marginally significant at the $p = 0.08$. The fact that the coefficients on "Salience × Other Treatment" and "Salience × Other Treatment × Bad Guy" have opposite signs relative to their self-condition counterparts (i.e. "Salience" and "Salience × Bad Guy") confirms that the salience of costume had an effect primarily in the "self" condition. The salience of costume had no significant effect in the "other" condition.

Drawing attention to "good guys" costumed identity causes them to lie more, while drawing attention to "bad guys" costumed identity causes them to lie less. This result is opposite to what we predicted.

3.4.3 Power Analysis

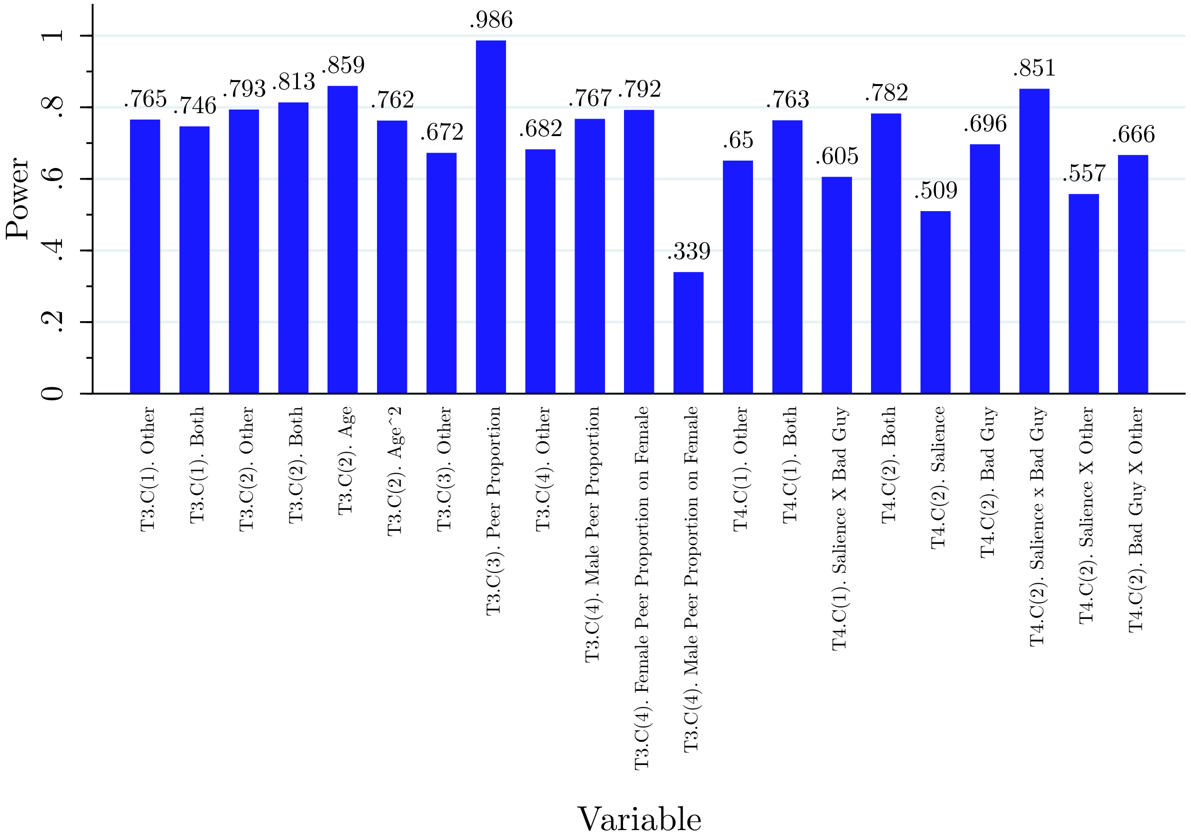
A potential concern of this analysis is that statistical tests may not be sufficiently well powered. We have 540 subjects placed into $2 \times 3 \times 2$ randomly assigned conditions. In addition, we test non-assigned features like age and peer effects. In many of our tests, statistical power would not appear to be problematic as the tests pool across conditions,[4] and appropriately so as the other conditions and variables of interest are orthogonal. However, for some analyses we estimate triple interactions leading to splits in the range of 1/8th of the sample (about 75 subjects) per cell. If statistical power is low, there is a risk that statistically significant results are lucky, generated from sampling variation and not from a true underlying effect.

We conduct an ex-post power analysis using Monte Carlo methods. We subject all of our results that are significant at the $p < 0.1$ level to this analysis. For each given logit regression we predict the probability that an individual will report a six. We then generate 1,000 data sets, each randomly assigning the outcome (reporting a six) to an observation based on the predicted probability. On each of these data sets we run the same regression and estimate the average marginal effects. The proportion of data sets that generated a significant coefficient at $p < 0.05$ is our estimate of statistical power for that coefficient. We conduct this analysis for every regression specification in the paper (four in Table 1 and two in Table 2; six total).

Figure 2 displays our estimated statistical power for each coefficient of interest that has $p < 0.1$ in the paper. Statistical power for most of our main results exceeds 0.7. The other condition reduces lying and we find that the power ranges from 0.65 to 0.793. The both

[4] For example, comparing beneficiaries tests between three conditions: self, other, and both. Since we pool the other orthogonal treatments within these three group, there are approximately 180 in each of these three conditions.

Figure 2: Reporting a Six as a Function of Age



Notes: "Tx.C(y). variable" refers to the coefficient of the variable found in Table x, Column (y). The values displayed in the figure show the statistical power of finding an effect of the size we estimated with critical value of $\alpha = 0.05$. This figure includes all tests for which the value was significant at $p < 0.1$. See main text for how the values are simulated.

condition ranges in power from 0.746 to 0.859. Our age results are presented in Col (2) and have a power of 0.859 and 0.762 on the linear and quadratic terms respectively. Turning to the peer effect in Col (3), we find that the test is powered at 0.986. When we decompose the peer effects by gender we find that the male peer effect on males is powered at 0.767 and the female effect on females is powered at 0.792.

In Table 4, our main specification is Col (2) and the two main variables of interest are Salience and Salience × Bad Guy, which are positive and negative respectively. These two coefficients are powered at 0.59 and 0.851 respectively. While there is no conventional threshold for statistical power in the same way there is for critical values, casual conversations amongst experimentalists suggest that a power of 0.8 is considered good.

3.5 Discussion

We found that our high-stakes manipulation had no effect. This is in-line with the results of previous studies (Mazar, Amir and Ariely, 2008; Fischbacher and Föllmi-Heusi, 2013). However, we ran a pilot in the preceding year with a similar design and slightly different stakes and found an effect. We view the result in this paper as a failed manipulation as we suspect that our high stakes were not sufficiently high. We believe that we cannot conclude much from this manipulation.

Changing the beneficiary had a large impact on lying. The "other" condition had much less lying than the "self" condition. Gino and Pierce (2010) and Cadsby, Du and Song (2016) show that people lie for others when they care for them, especially if the other person is part of their group. Michailidou and Rotondi (2019) find that individuals are not willing to lie to benefit an out-group member and that lying for an in-group member is uncommon. Our results are consistent with these findings as the beneficiary in our "other" condition is an anonymous stranger. We also find that the "both" condition reduces lying for one's own gain but not for the gain of others. Wiltermuth (2011) and Gino, Ayal and Ariely (2013) find that people are more likely to view dishonesty as morally acceptable when their dishonesty would benefit others. The past studies found that children lie more for both themselves and others when there is an opportunity to lie to help others. Our results are contrary to this finding. Instead, honesty for reporting the “other” die spilled over to reporting the “self” die but not vice versa. This could be caused for a number of reasons. The cognitive load placed on the children for engaging in the two activities may have caused them to lie less. Or the inconsistency between lying about one report but not the other may have made rationalizing lying more difficult. A third potential explanation is that the children may have thought they would be more likely caught lying if they reported a six for themselves and not a six for “other”. We cannot test between these different hypotheses unfortunately.

We find no direct effect of gender. This is consistent with the result of Nieken and Dato (2016). However, some research suggests that men are more likely to lie to advance themselves, while women are more likely to lie to advance others (DePaulo et al., 1996; Feldman et al., 2002; Erat and Gneezy, 2012b; Dreber and Johannesson, 2008b). The sample here differs from other studies as it is primarily children.

It is surprising to us that costume salience led to behavior incongruous with one’s costume. Past research found the salience of identity to have effects congruous with the identity (Cohn, Fehr and Maréchal, 2014; Cohn, Maréchal and Noll, 2015a; Celse and Chang, 2017). We offer some speculation regarding this finding. One possibility is that the costume led to congruous ethical behavior earlier in the day and this bolstered a moral licensing/self-conscious effect later during our experiment. However, for this explanation to make sense it would require that the salience manipulation not only made the ethical orientation of one’s costume more salient but also made past recent ethical actions more salient. While possible, the manipulation made no mention of recent activity.

An alternative explanation is that the salience manipulation felt like an announcement or confession of moral disposition to an *observing* adult. Past research has shown that people conform more to social norms when they feel like they are being observed even though they are not actually being observed. For example, Haley and Fessler (2005) found that the image of stylized eyespots increased giving in a dictator game. Mol, van der Heijden and Potters (2020) showed that subjects in a virtual reality environment were less likely to cheat when a virtual observer was watching, compared to when the virtual observer was looking at a smartphone. If the children thought that the adults were monitoring them, publicly announcing to an adult that they are "good guys" may have reduced the feeling of being observed, while announcing to an adult that they are "bad guys" may have increased the feeling of being observed.

There are a couple of reasons we believe this explanation is less convincing than the moral licensing explanation. First, several papers find no effect of watching eye cues (Pfattheicher, Schindler and Nockur, 2019; Ayal, Celse and Hochman, 2019), suggesting that the effect may not be very robust. Second, and more importantly, the experimenters who asked the priming question were different from the experimenters who ran the game and gave the candy. The experimenters who asked the priming question were also physically distant, at about 25 feet away from the point at which the trick-or-treaters reported their die rolls. This means that for this "feeling of differential monitoring" hypothesis to be true, trick-or-treaters would have to feel that their communication to one experimenter affected other experimenters 25 feet away without the use of communication. This sounds less plausible to us but it remains a possibility. An alternative design that makes the salience manipulation private could potentially disentangle this effect.[5]

While modern practice has mostly done away with the norm-violating "tricks" of the past, in our contrived setting, victimless lying for candy is pervasive. And like the paired practice of costume-wearing and norm violations in past traditions, we see a relationship

[5] We hoped to conduct a follow-up experiment in which trick-or-treaters would be primed with the same question, "Is (blank) a good guy or a bad guy?", but would not need to report it aloud. Unfortunately, the global pandemic of 2020 interfered with the ability to conduct follow-up Halloween experiments.

between costume and lying in our study. “Bad guys” in the no-salience condition lie more than “good guys”. This suggests that there is either a causal relationship or a self-selection effect of wearing a “bad guy” costume. Even if the effect is entirely through self-selection, this is still an interesting relationship. It means that those prone to norm violations select a consistent costumed identity.

Identity plays a second role in our study in the form of gender identity. We find that reporting a six increases the probability that a child later in line reports a six. This effect is only significant within gender. We offer a couple of reasons why this may be the case. First, if children tend to befriend within gender, trick-or-treat with friends, and emulate their friends, this could generate the result. Though we did not keep records, our impression was that most groups of trick-or-treaters were not groups of friends but families. This seems to be especially true for the younger trick-or-treaters, though there were certainly some groups of friends. However, the great majority of the trick-or-treaters within a participant’s group of ten would have been strangers even if they had approached the house with family or friends. Indeed, many participated without any other individuals from an observable group, and those that participated with someone from their group usually only had no more than one or two accompanying them. Our sense is that the effect of friend-emulation would have to be exceptionally strong to be the sole-driver of our within-gender peer effect. We offer another explanation for the within-gender peer effect: because socially acceptable behavior is often gender-specific, preferentially emulating others within one’s own gender is a simple and effective heuristic. Participants may have applied such a heuristic even when the norm does not vary by gender, as is the case for lying.

While we suggest caution in over-inferring from our boundary-case context (i.e. extreme costumes for children), the implications of our results suggest that policies that require particular kinds of dress may influence the behavior of the wearers. This has been a motivation for educators to adopt school uniforms and indeed Evans, Kremer and Ngatia (2008) found that school uniforms in Kenya reduced absenteeism. We speculate that the effect of clothing on an individual's sense of identity may also be a motivation for business dress codes. Wearing a suit serves to signal professionalism to clients, but in many businesses employees wear suits even on days for which they do not meet with clients. This suggests that such dress codes also serve an internal purpose. As a counterpoint, explicit "casual Friday" policies encourage less formal clothing as it is perceived to create a more convivial work environment.

Discussion Points: 1. Why did we get moral licensing instead of behavior in congruity with clothes? a) Maybe they did congruent things before. b) Maybe because it was a public announcement of identity to a watching adult. 2. Age, does candy matter as much? 3. Interesting that the selection effect goes in opposite direction of priming effect.

One pivotal finding from this study is the incongruent nature of lying outcomes under the priming condition. Our a priori hypothesis proposed trick-or-treaters will be invested in their costume choice and subsequently a reinfornced priming of such costume will induce actions congruent with their costume's theme:good guy or bad guy. More contextually, we expected individuals dressed as "good guys" would lie less when given the opportunity to lie for greater returns, candies. In actuality, our results indicate that primed trick-or-treaters dressed as "bad guys" lied less for candy compared to their "good guy" counterparts. [This is not correct]. A proposed explanation to this observed phenomena is moral licensing: "good guys" feel their costume equates to moral tokens providing them less incentive to follow norms and more incentive to lie making use of their good deed tokens. Conversely, given that "bad guys" have now been observed as category "bad" by the experimenter conducting the primer, they may feel persuaded to lie less in order to reinforce and overtly show their true identity does not align with their "bad guy" costume. When comparing priming vs. non-priming groups, the data indicates that "good guys" lying for more candies is more frequent for primed conditions versus non-primed conditions.

One of the intriguing finding that we get is the counter-expecting priming effect. Intuitively, after we prime the participants to focus on their costume, they would be more invest in their character role playing - expecting lying less for good guys and lying more for bad guys. This is also the mechanism behind in custom, identity, and norm violation behavior in our theoretical context. However, the results turn out to be opposite. This does not mean the link has been torn down, but maybe just two sides of the same story. The rationale we have here is moral licensing: the good guys lie more since they get moral licensed. Yet still there might exist some "moderator" that we overlooked. One of the differences between laboratory experiment and field experiment is "context". The context here is that we take advantage of Halloween costume and trick or treat tradition to delve in act in identity and norm violation, which a little mischievous behavior is in fact following alone with social expectation. Maybe integrity good guy concept obsolete (Yes, Iron man rock compares to the Captain); Maybe good guys' player suppose they should get bonus candy for the effort of saving world. This is the part when moral licensing jumps in, because good guys lie less for themselves without priming but it increases for the group that primed. Nevertheless, moral licensing says little about bad guys behavior. We believe some characters to those kids may not be all bad, which means when chances come, they might want to clear that stigma for them. Cosplay is about role-playing and honoring based on passion to the character. There is possibility that through priming, we trigger the acting part of them to extend the story of their character. The most far-reaching discussion would related to why these children chose these costume and if this related to their understanding to social expectation, social norm, and norm violation behavior, but for here, we attribute this beautiful cross-over interaction to Halloween trick or treat context.

Several sample selection and experimental issues are important to note. First, there might exist selection bias since the costumes were pre-selected by the trick-or-treaters. Wearing a "bad guy" costume increases the probability of reporting a six for one's self. It is possible that wearing a "bad guy" costume affects wearer's lying behavior. But It is also possible that a person who is more likely to lie is also more likely to chose a "bad guy" costume. However, this does not pose a problem for inferring the effect of priming since priming was randomized in our design. It would be interesting to see whether the result will be different, if the clothes are given to them, instead of wearing the pre-selected costumes.

Naturally, as with any empirical exercise, questions of external validity remain. In particular, whether results of this experiment were particular to the context of trick-or-treater during Halloween, whether candy provides sufficient incentive. For instance, we believe how many candies you get during Halloween night matters to a 8 years old kid, but how much it matters to a 15 years old teenager. It is possible that lying declines after age of 12 or 13 due to less motivation of getting more candies.

Another concern is that the moral licensing effect of good guys might be caused by the prior good deeds that they did while wearing the "good guys" costume before coming to the trick-or-treating event on the Halloween night. More research is needed to find the reason why we found moral licensing effect instead of behavior in congruity with clothes.

Interestingly, in the priming condition, "bad guys" lied less compared to the un-primed "bad guys". This result might be caused by the public announcement of "bad guys" identity. When subjects reported "bad guys" to us, they might be afraid to lie since "bad guys" have the tendency to be watched or judged. It might be a different result if we priming them without reporting "good or bad" to us.

Last, due to the complexity of the design where the kids need to roll two dice and reported twice for both treatment, the decrease in the probability of lying can be caused by the heavy cognitive loading for the kids. Thus, it would also be interesting to study whether their lying behavior changes if they roll only one die that determines the payoff for both themselves and others, instead of one die for each.

3.6 Conclusion

We study the impact of Halloween costumes on the ethical behavior of 544 trick-or-treaters. We find that lying is more common when oneself is the beneficiary than when someone else is the beneficiary. We find that having the opportunity to lie for both oneself and someone else causes more honesty when reporting for oneself. Lying peaks at age 12 and is influenced by the lying of peers. In particular, peer effects are strong and predominantly within gender. Finally, we find that rendering the ethical orientation of one's costume salient leads to more or less lying depending on whether one's costume is a good guy or a bad guy. The salience of costume causes "good guys" to lie more and "bad guys" to lie less, consistent with a moral licensing. Though we believe this is the more plausible explanation, we cannot falsify the alternative explanation that trick-or-treaters believed that publicly reporting one's identity would cause differential monitoring from the experimenters.

This paper extends the literature on lying, and it is one of the first papers in economics that connects clothing and identity to ethical behavior. By leveraging a naturally-occurring cultural tradition — costume-wearing on Halloween — we elucidate the relationship between clothing and behavior using an extreme boundary case. The results also inform us on the nature of costume-wearing itself, implying that the tradition may have served as a means to encourage norm-violations by temporarily changing people's sense of identity.

Appendix

Appendix A (Chapter 1)

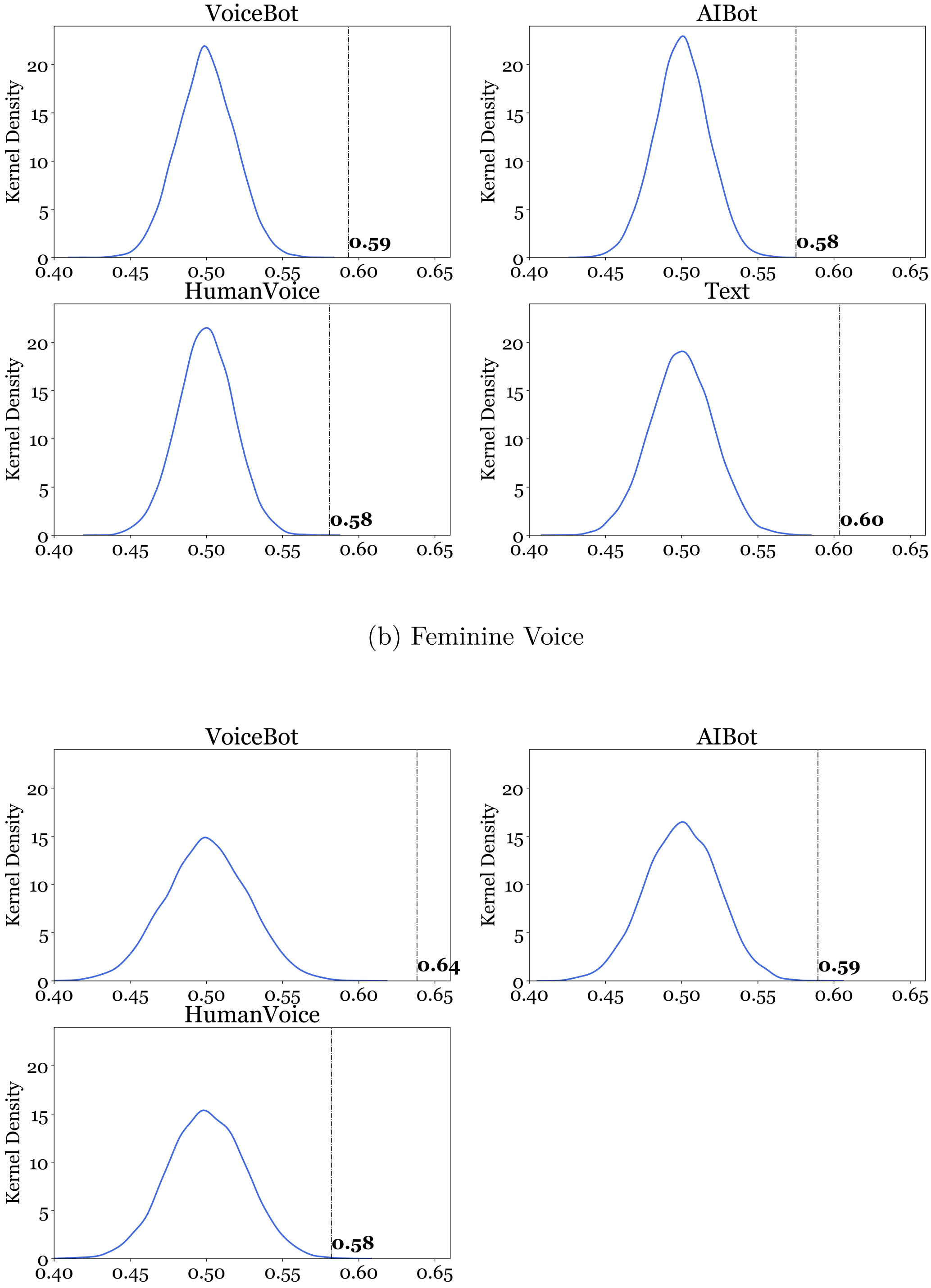
A. Simulation

Empirical distributions of successful coin tosses may deviate from their theoretical counterpart (i.e., binomial distribution), even if the coins are fair and everyone reports their outcomes truthfully. Due to random fluctuations, actual frequencies of successful coin flips may not exactly match the expected frequencies. In this section, we explore using simulations whether the observed treatment effects can, in principle, be explained by random fluctuations. To this end, we simulated 10,000 coin flipping experiments for each treatment with the same number of subjects and coin flips as in the respective treatments. In the simulations, we assume that each coin toss is generated by a binomial process with an underlying success rate of 50% (i.e., truthful reporting).

Figure A1 in the Appendix shows the share of successful coin tosses compared to our simulation. In order to get a clear comparison with the real coin-toss probability, we simulated each of the treatments 10,000 times by using the same sample size for each treatment. Each sub-figure represents the distribution for the probability of flipping a head per each treatment. From the four sub-figures, the probability of reporting a head (or 1 in this case, while tail is 0.) is significantly higher than the maximum probable range of a fair share. This indicates that in all of our treatments, there is a prevalence of lying that is significantly higher than fair share tosses.

Figure A1b shows a similar setup of comparing successful coin tosses focusing in a feminine voice. Although all the sub-figures show only suggestive evidence of lying in all of our treatments, however, there is a fair convincing evidence that the probability of head is larger when reporting to a non-human, especially to a voicebot. Lastly, Figure A1c shows the share of successful coin tosses for a masculine voice. Interestingly, when reporting to a masculine voice, the probability of reporting a head is fairly smaller for voicebot, indicating that the voice gender in voicebot seems to influence participants' perspective of self-honesty that subsequently influences their propensity of lying.

Figure A1: Share of Successful Coin Tosses (Actual vs. Simulation)



(a) All Sample

(b) Feminine Voice

(c) Masculine Voice

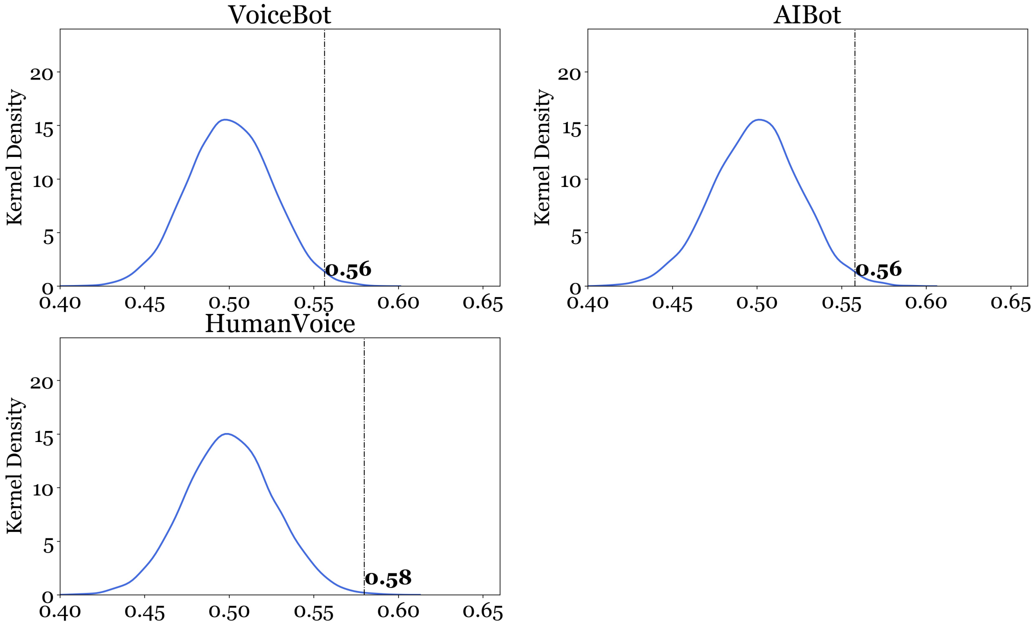


Figure A2: Total Heads by Treatments

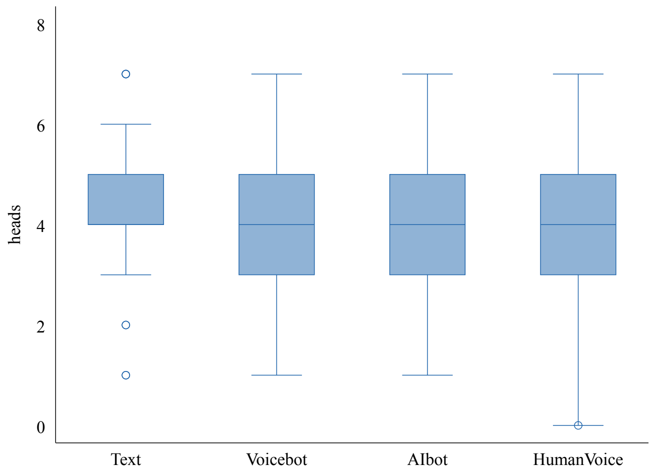
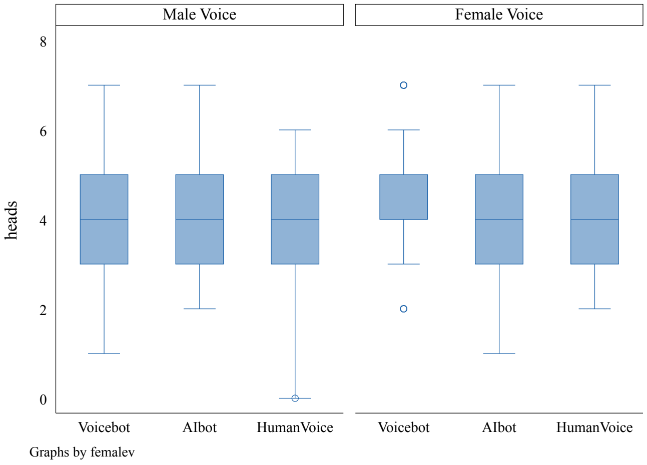


Figure A3: Total Heads by Treatments and Voice Gender



B. Table and Figure

Table A1: Summary Statistics - Demographics

	All	**Text**	**VoiceBot**	**AIBot**	**HumanVoice**	**p-value**
Female	0.521	0.476	0.5	0.56	0.533	0.65
	(0.5)	(0.5)	(0.5)	(0.497)	(0.499)	
College	0.531	0.56	0.51	0.534	0.524	0.921
	(0.499)	(0.497)	(0.5)	(0.499)	(0.5)	
White	0.809	0.798	0.817	0.793	0.829	0.904
	(0.393)	(0.402)	(0.387)	(0.405)	(0.377)	
Age Category	3.667	3.25	3.558	3.802	3.962	0.073
	(1.561)	(1.551)	(1.538)	(1.637)	(1.421)	
Feminine Voice	0.494		0.452	0.543	0.514	0.2
	(0.5)		(0.498)	(0.498)	(0.5)	
Subjects	409	84	104	116	105	

Note: This table reports means and standard deviations (in parentheses) of participants' demographic information. The last column contains p-values for the null hypothesis of perfect randomization (χ^2-tests for all variables).

Table A2: Summary Statistics - Toss by Participant Gender and Voice Gender

	All	**VoiceBot**	**AIBot**	**HumanVoice**	**p-value**
Toss - Fem. x Fem. Voice	0.56	0.548	0.548	0.571	0.835
	(0.50)	(0.50)	(0.50)	(0.50)	
Toss - Male x Fem. Voice	0.666	0.733	0.646	0.603	0.055
	(0.47)	(0.44)	(0.48)	(0.49)	
t-test p-value	0.000	0.000	0.039	0.557	
Toss - Fem. x Masc. Voice	0.536	0.52	0.532	0.579	0.552
	(0.50)	(0.50)	(0.50)	0.50)	
Toss - Male x Masc. Voice	0.589	0.591	0.589	0.581	0.974
	(0.49)	(0.49)	(0.49)	(0.49)	
t-test p-value	0.113	0.156	0.270	0.969	

Table A3: Factor Loadings and KMO: Chatbot & AI Attitude

	AI 1	AI 2	AI 3	Unexplained	KMO
	ChatbotPositive	AINegative	AIFamiliarity		
I use chatbots very often.				0.49	0.76
I believe chatbots understand me well.	0.45			0.38	0.82
I feel chatbots can solve my problems.	0.49			0.31	0.81
I feel frustrated when I am interacting with chatbots.	-0.44			0.48	0.86
I rather interact with a chatbot than a person.	0.48			0.47	0.84
I am familiar with AI.			0.7	0.32	0.7
I trust AI.				0.31	0.82
I fear AI.		0.54		0.32	0.74
AI will destroy humankind.		0.57		0.26	0.74
AI will benefit humankind.				0.43	0.85
AI will cause many job losses.		0.43		0.48	0.87
				Overall	0.8

Table A4: Factor Loadings and KMO: Trust and Lying Attitude

	Trust 1	Trust 2	Trust 3	Unexplained	KMO
	Honesty	Trusting	Trustworthiness		
People can be trusted.		0.58		0.15	0.83
People are helpful.		0.59		0.1	0.78
People are fair.		0.56		0.12	0.8
I am trustworthy.			0.7	0.13	0.72
I am fair.			0.71	0.13	0.72
I never tell lies to family/friends.	0.5			0.13	0.82
I never tell lies to my partner/spouse.	0.47			0.25	0.85
I never tell lies to acquaintances.	0.52			0.11	0.78
I never tell lies to strangers.	0.51			0.21	0.8
				Overall	0.79

Table A5: Factor Loadings and KMO: Feminine Voice Attitude

	Fem. Voice 1 Attractive	Fem. Voice 2 Trutworthy	Fem. Voice 3 Authoritative	Unexplained	KMO
Attractiveness	0.65			0.14	0.94
Authoritativeness			0.74	0.1	0.87
Confidence			0.64	0.13	0.89
Familiarity	0.56			0.22	0.96
Likeability		0.5		0.13	0.87
Trustworthiness		0.71		0.13	0.93
Warmth				0.19	0.91
				Overall	0.91

Table A6: Factor Loadings and KMO: Masculine Voice Attitude

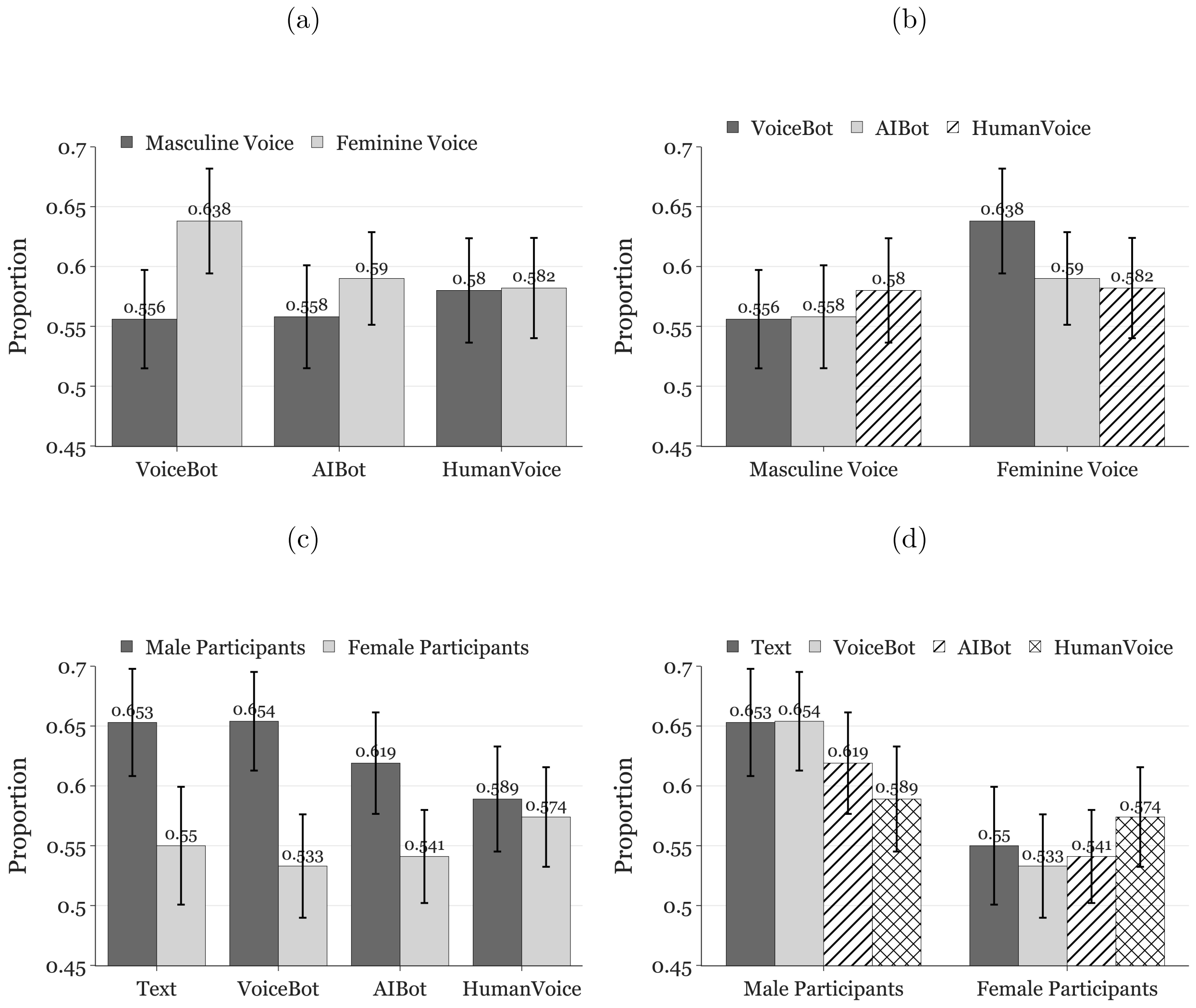
	Masc. Voice 1 Attractive	Masc. Voice 2 Trutworthy	Masc. Voice 3 Authoritative	Unexplained	KMO
Attractiveness	0.41			0.28	0.94
Authoritativeness			0.78	0.1	0.87
Confidence			0.56	0.15	0.89
Familiarity	0.75			0.18	0.96
Likeability		0.45		0.13	0.87
Trustworthiness		0.78		0.11	0.93
Warmth	0.44			0.19	0.91
				Overall	0.91

Table A7: The Gender Effect of Participants

	(1)	(2)	(3)	(4)
	Text	**VoiceBot**	**AIBot**	**HumanVoice**
Female	-0.086**	-0.151***	-0.116***	0.004
	(0.035)	(0.036)	(0.033)	(0.042)
Subjects	83	102	115	103
Observations	581	714	805	721

The table shows the average marginal effects. Demographic and attitudinal variables (i.e., chatbot and AI, trust and lying) are controlled in the regression. Clustered standard errors are in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

Figure A4: Propensity of Lying


(a)

(b)

(c)

(d)

Note: percentage of coin tosses reported as heads are shown above each bar.

C. Factor Analysis

Our survey included a large number of questions regarding participants' attitudes toward Chatbot and AI, their perspectives about the feminine and masculine voicebot's voice, as well as questions about individual trust and lying characteristics (see Table A3 - A6 in the Appendix). We used factor analysis (Basilevsky, 2009) to reduce the large number of variables into fewer number of factors by combining the relevant information. We performed factor analysis procedures following Yong et al. (2013) to select the variables that may share latent variables and also to check for the proportion of variance in our variables that might be caused by underlying factors (sample adequacy).[6] We applied the method of principal component factors to extract latent factors. The Varimax rotation procedure was used to obtain the factor loadings. To determine how many factors to retain, we used eigenvalues benchmark with a cut-off of 1 in conjunction with scree test.[7] Using an eigenvalue cut-off of 1, we retained several set of factors that explained a variance of 79% to 90% of each selected group. Table A3 - A6 present the factor loadings of each variable on these factors after the Varimax rotation.

The rotated factor loadings in principal component factor tables show that our factors are fairly desirable with at least 2 variables per factor that have loading values above 0.32.[8] The factor loadings are the correlation coefficients between the indicators (rows) and factors

[6] First, we checked if there was a patterned relationship among our variables using correlation matrix. Variables that had a correlation coefficient above 0.3 or below -0.3 with at least 1 variable within their groups were selected for factor analysis. Then, we performed Bartlett's Test of Sphericity to confirm that our sample had patterned relationships. The test confirmed patterned relationships among the selected variables ($p<0.001$). Finally, we determined if our sample was suitable for factor analysis by looking at the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy with a cut-off value above 0.5 as acceptable threshold. The KMO values were all above 0.5 and ranging from 0.79 to 0.91.

[7] A rule of thumb, Kaiser's criterion, suggests retaining all factors that are above the eigenvalue of 1 (Kaiser, 1960). However, this criterion may result in overestimation in the number of factors extracted. Therefore, we also used the scree test in conjunction with the eigenvalues to determine the appropriate number of factors to retain.

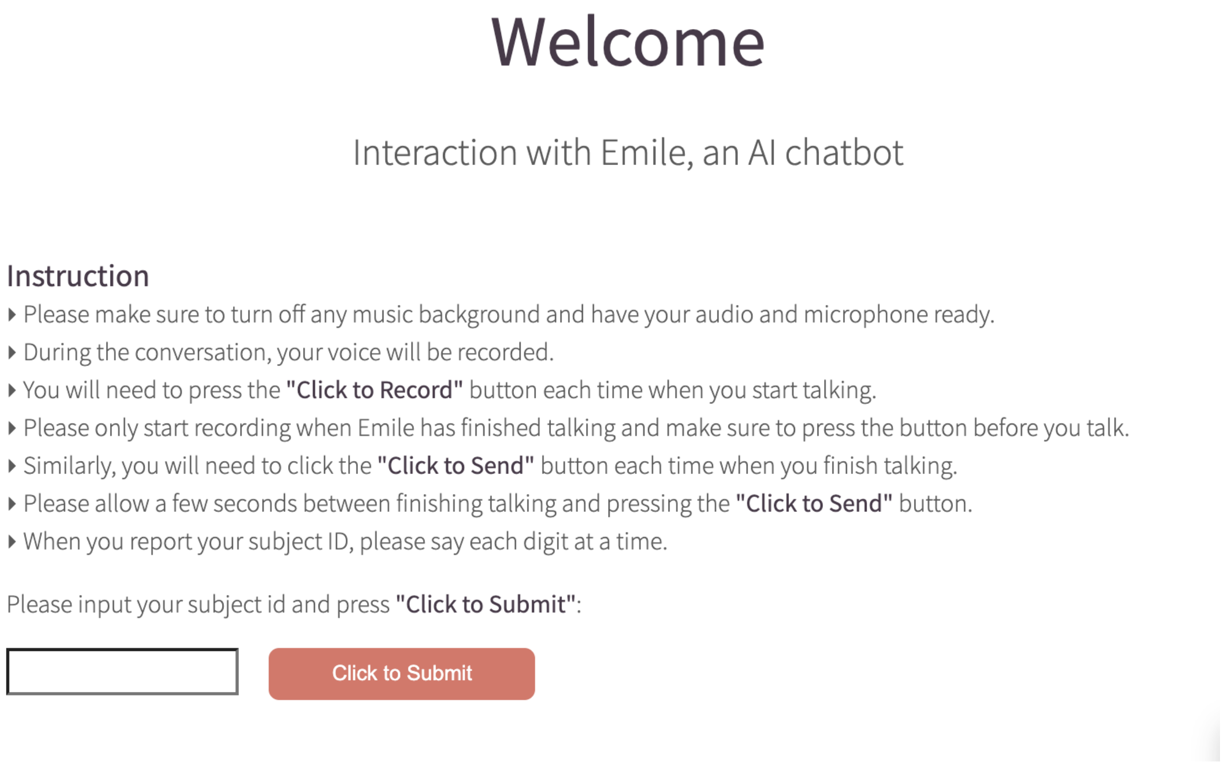
[8] Using an alpha level of 0.01 (two-tailed), a rotated factor loading for a sample size of at least 300 would need to be at least 0.32 to be considered statistically meaningful (Tabachnick et al., 2007). A factor loading of 0.32 gives approximately 10% of the overlapping variance.

(columns) and provide the basis for interpreting the different factors. High factor loading indicate the dimensions of the factors are better accounted for by the variables, as a factor loading for a variable is a measure of how much the variable contributes to the factor.

Table A3 - A6 show that our variables can be grouped into three factors per each category. Trust and Lying Attitude has 3 factors which are related to participants self-rated lying attitude (Honesty), attitude toward other people including trust, fairness and helpful (Trusting), and self-sense of trustworthiness and fairness (Trustworthiness). Chatbot and AI attitude can be grouped into 3 factors which are the positive attitude toward chatbot (ChatbotPositive), the negative attitude toward AI (AINegative), and the familiarity with AI (AIFamiliarity). Our masculine and feminine voice attitude tend to have the same factor groupings which are related to attractiveness and familiarity (Attractive), likeability and trustworthiness (Trustworhy), and authorativeness and confidence (Authorative). For our regression analysis, we mainly utilized the factor scores of the three factors for Trust and Lying attitude, Chatbot and AI Attitude, and Masculine and Feminine Voice attitude as controls when indicated.

D. Screenshots of Online Voice Reporting Interface

Figure A5



Welcome

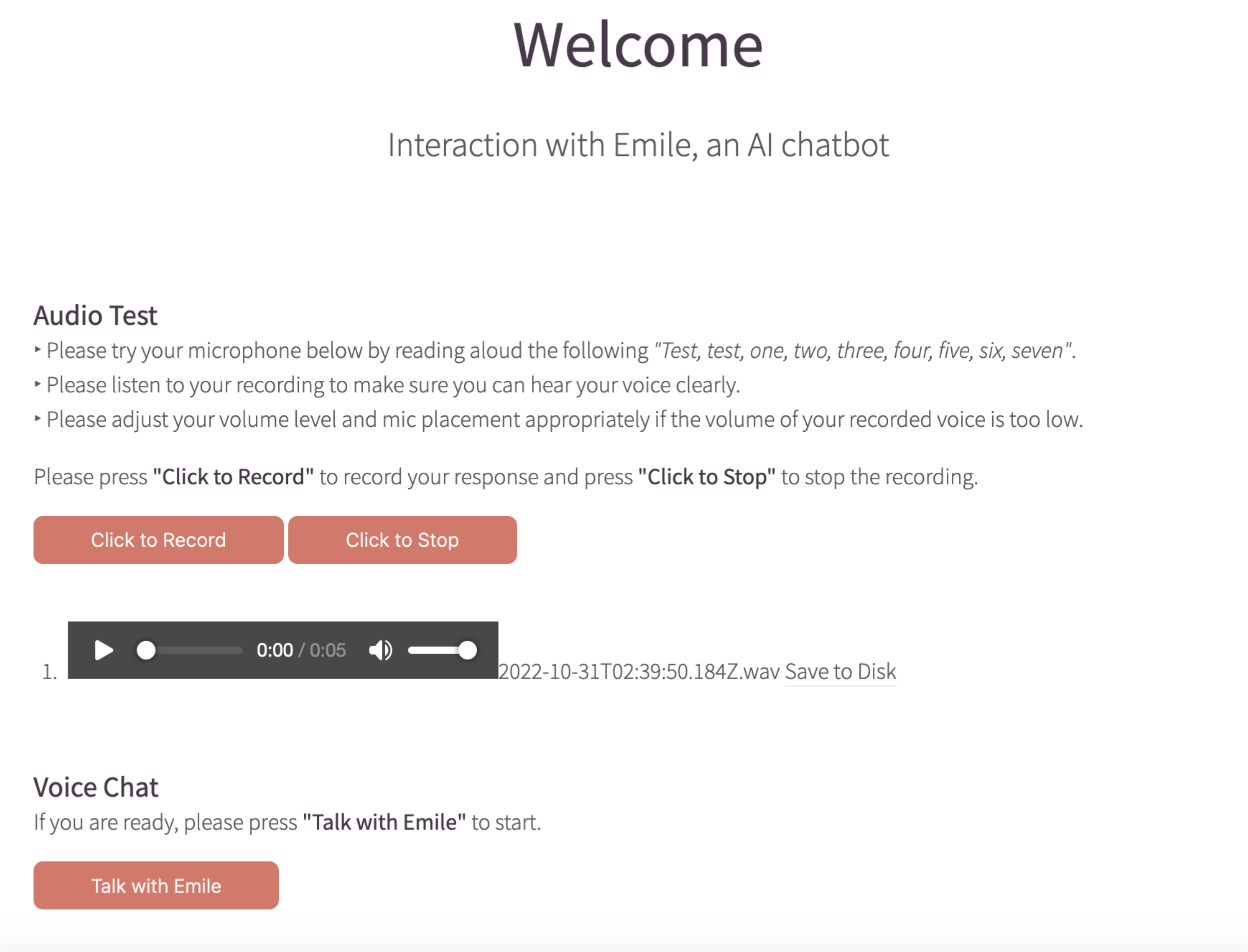
Interaction with Emile, an AI chatbot

Instruction

- Please make sure to turn off any music background and have your audio and microphone ready.
- During the conversation, your voice will be recorded.
- You will need to press the **"Click to Record"** button each time when you start talking.
- Please only start recording when Emile has finished talking and make sure to press the button before you talk.
- Similarly, you will need to click the **"Click to Send"** button each time when you finish talking.
- Please allow a few seconds between finishing talking and pressing the **"Click to Send"** button.
- When you report your subject ID, please say each digit at a time.

Please input your subject id and press **"Click to Submit"**:

Figure A6



Welcome

Interaction with Emile, an AI chatbot

Audio Test

- Please try your microphone below by reading aloud the following *"Test, test, one, two, three, four, five, six, seven"*.
- Please listen to your recording to make sure you can hear your voice clearly.
- Please adjust your volume level and mic placement appropriately if the volume of your recorded voice is too low.

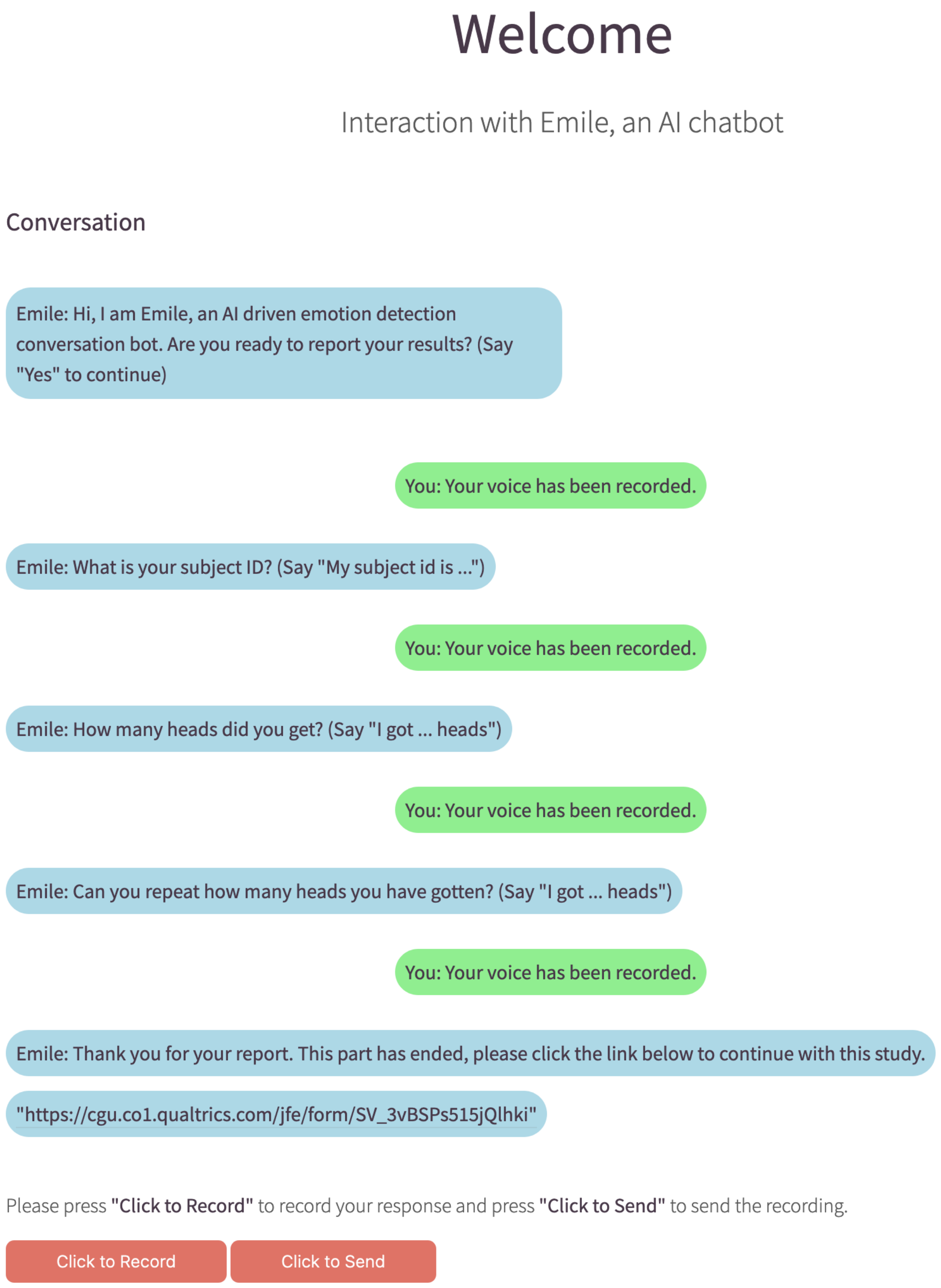
Please press **"Click to Record"** to record your response and press **"Click to Stop"** to stop the recording.

1. 2022-10-31T02:39:50.184Z.wav Save to Disk

Voice Chat

If you are ready, please press **"Talk with Emile"** to start.

Figure A7



Welcome

Interaction with Emile, an AI chatbot

Conversation

Emile: Hi, I am Emile, an AI driven emotion detection conversation bot. Are you ready to report your results? (Say "Yes" to continue)

You: Your voice has been recorded.

Emile: What is your subject ID? (Say "My subject id is ...")

You: Your voice has been recorded.

Emile: How many heads did you get? (Say "I got ... heads")

You: Your voice has been recorded.

Emile: Can you repeat how many heads you have gotten? (Say "I got ... heads")

You: Your voice has been recorded.

Emile: Thank you for your report. This part has ended, please click the link below to continue with this study.

"https://cgu.co1.qualtrics.com/jfe/form/SV_3vBSPs515jQlhki"

Please press **"Click to Record"** to record your response and press **"Click to Send"** to send the recording.

Appendix B (Chapter 2)

A. Table and Figure

Table B1: Accoustic Features

Feature	**Description**	**Mean (Std.)**	**Min-Max**
Pitch	Fundamental Frequencies, via f0 frequencies (range: 75-600 Hz).	200.42 (51.09)	83.06-359.07
Intonation	Std. Deviation of Fundamental Frequencies, a measure of pitch variation.	73.09 (34.65)	8.01-175.59
Roughness	Harmonic-to-Noise Ratio (HNR), a measure of roughness, via the forward cross-correlation method (mean value; time step=0.01 s; min pitch=75 hz).	8.10 (2.35)	1.49- 13.98
Voice Perturbation: Shimmer	A measure of amplitude variation, via the Amplitude Perturbation Quotient (APQ3) measuring the average absolute difference between a period amplitude and the average of amplitude of the neighbors, divided by the average (shortest period=0.0001 s; longest period=0.02 s; max period factor=1.3; max. amplitude factor=1.6).	0.07 (0.02)	0.03-0.13
Voice Perturbation: Jitter	A measure of local pitch variation, via Relative Average Perturbation (RAP) measuring the average absolute difference between a period and the average of that period and the two neighbors (shortest period=0.0001 s; longest period=0.02 s; max. period factor=1.3).	0.02 (0.01)	0.00-0.04
Formants	The formant is the broad spectral maximum that results from an acoustic resonance of the human vocal tract. It is a characteristic of the resonances of the space and a concentration of acoustic energy around a particular frequency in the speech wave. Each formant corresponds to a resonance in the vocal tract. The formant with the lowest frequency is called F1, the second F2, the third F3, and the fourth F4.		

Table B2: Summary Statistics: Cooperation vs. Deception

	Golden Balls (n=236)				**Real-life Trial (n=108)**			
Variable	Female		Male		Female		Male	
	Cooperative	Deceptive	Cooperative	Deceptive	Truthful	Deceptive	Truthful	Deceptive
Pitch	227.41	228.35	212.81	209.23	179.56	185.37	115.34	125.77
	(31.74)	(26.93)	(52.51)	(53.79)	(22.44)	(25.87)	(24.08)	(22.5)
Intonation	73.97	73.06	108.15	108.86	43.06	37.53	31.75	36.43
	(20.67)	(15.9)	(24.61)	(22.64)	(12.52)	(12.04)	(19.01)	(24.95)
HNR	8.57	8.77	7	6.78	9.71	9.05	8.25	6.28
	(2.13)	(2.3)	(1.58)	(2.4)	(2.4)	(2.17)	(2.34)	(2.18)
Shimmer	0.065	0.065	0.085	0.088	0.065	0.066	0.063	0.077
	(0.016)	(0.016)	(0.018)	(0.02)	(0.02)	(0.017)	(0.016)	(0.026)
Jitter	0.014	0.013	0.02	0.022	0.013	0.014	0.013	0.015
	(0.006)	(0.005)	(0.008)	(0.008)	(0.003)	(0.006)	(0.006)	(0.008)
N	74 (58%)	53 (42%)	58 (53%)	51 (47%)	21 (32%)	44 (68%)	32 (74%)	11 (26%)

Table B3: Machine Learning Model: Full Features

Model	Accuracy	SD	Model	Accuracy	SD
Golden Balls: Female			***Golden Balls: Male***		
Naive Guess	0.502	0.082	Naive Guess	0.502	0.092
Decision Tree	0.566	0.084	Decision Tree	0.59	0.09
Random Forest	0.599	0.079	Random Forest	0.647	0.081
Real-life Trial: Female			***Real-life Trial: Male***		
Naive Guess	0.499	0.101	Naive Guess	0.502	0.123
Decision Tree	0.708	0.11	Decision Tree	0.683	0.118
Random Forest	0.747	0.09	Random Forest	0.79	0.105

Notes: The naive guess shows the initial probability of correctly predicting variable of interests. All procedures are repeated 1,000 times using random state sampling with Synthetic Minority Oversampling Technique (SMOTE). The accuracy is an average prediction accuracy after 1,000 iteration. We use 75% training set and 25% test set. Full features include voice features, emotion from voice, text sentiment, and additional features such as uni-gram text features (only for GB).

Table B4: Cross-context Validation: Full Features

Model	Accuracy	SD	Model	Accuracy	SD
GB to RT: Female			***GB to RT: Male***		
Naive Guess	0.498	0.056	Naive Guess	0.505	0.067
Decision Tree	0.546	0.065	Decision Tree	0.449	0.076
Random Forest	0.569	0.052	Random Forest	0.412	0.059
RT to GB: Female			***RT to GB: Male***		
Naive Guess	0.498	0.045	Naive Guess	0.50	0.048
Decision Tree	0.516	0.044	Decision Tree	0.498	0.044
Random Forest	0.524	0.038	Random Forest	0.50	0.034

Notes: The naive guess shows the initial probability of correctly predicting variable of interests. All procedures are repeated 1,000 times using random state sampling with Synthetic Minority Oversampling Technique (SMOTE). The accuracy is an average prediction accuracy after 1,000 iteration. We use 90% of the source and target set per iteration. Full features include voice features, emotion from voice, and text sentiment.

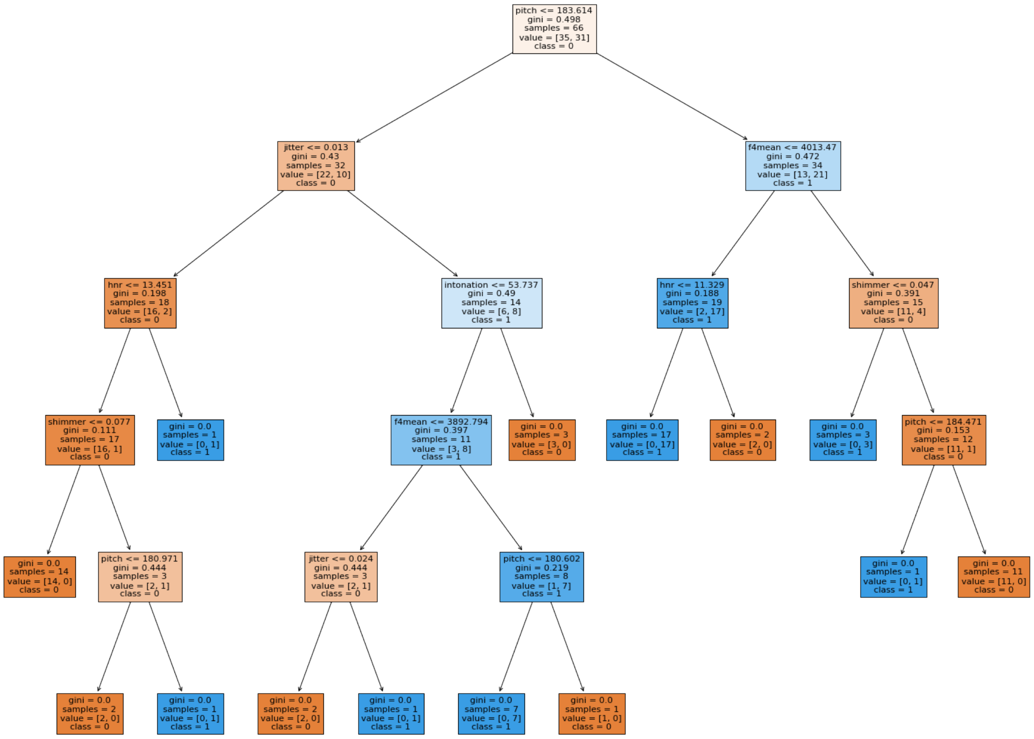
Table B5: Pair-wise Correlation Coefficients for Emotions from Voice

Variable	Calm	Anger	Joy	Sorrow	Energy
Calm	1.0				
Anger	-0.071	1.0			
Joy	-0.592*	-0.248*	1.0		
Sorrow	-0.397*	-0.007	-0.443*	1.0	
Energy	-0.501*	-0.028	0.917*	-0.529*	1.0

Note: Significance: *p<0.05

Notes: Significance: *p<0.05.

Figure B1: Decision Tree Example



B. Machine Learning Methods

Synthetic Minority Over-sampling Technique (SMOTE)

Imbalanced classes (or groups) can result in severe ML prediction bias (see Jacobusse and Veenman (2016)). Therefore, we need to correct this by method of resampling (e.g., over-sampling, under-sampling, etc.). In this paper, we use Synthetic Minority Over-sampling Technique (SMOTE) to obtain a more efficient way of over-sampling (see Chawla et al. (2002)). SMOTE works by creating synthetic samples from the class minority instead of creating copies of records. The algorithm selects two or more similar instances by using a distance measure and perturbing an instance one attribute at a time by a random amount within the difference to the neighboring instances. Its theory bases that the feature space of minority class instances is similar (Zheng et al., 2015). For each instance x_i in minority class, SMOTE searches its k nearest neighbors and one neighbor is randomly selected as x_0 (we call instances x_i and x_0 seed sample). Then a random number between [0, 1] δ is generated. The new artificial sample x_n is created following the formula shown in Equation B1:

$$x_n \; = \; x_i \; + (x_0 - x_i) \; \times \; \delta \tag{B1}$$

Compared with random oversampling method, SMOTE can effectively avoid the problem of over-fitting of classifiers without removing samples or adding replications.

Decision Tree

We use decision tree as our basic machine learning algorithm to classify two classes. Decision tree is a non-parametric supervised learning model for classification that can be use for both categorical and continuous variables. The tree-structure diagram of decision tree model helps us to visualize the model and interpret the results easily. The hierarchical structure of a decision tree leads us to the final outcome by traversing through the nodes of the tree. Each node consists of a feature which is further split into more nodes. In our Decision Tree model, we use Gini index as the splitting measures. Gini index measures the degree or probability of a particular variable being wrongly classified to one of the classes when it is randomly chosen. The degree of Gini index varies between 0 and 1, where 0 denotes that all elements belong to a certain class or if there exists only one class, while 1 denotes that the elements are randomly distributed across various classes. A Gini index of 0.5 denotes equally distributed elements into some classes. Our Decision Tree algorithm will first decide one predictor that is able to split the outcomes with the least Gini index to be the first node. In the event when the predictors are continuous variable, the algorithm will try to find the appropriate slice of one of the predictors that yield the least Gini index. This procedure is repeated to create subsequent nodes until all the outcomes are classified correctly. The algorithm will follow the formula shown in Equation B2:

$$Gini \ = \ 1 \ - \sum_{i=1}^{n} p_i^2 \tag{B2}$$

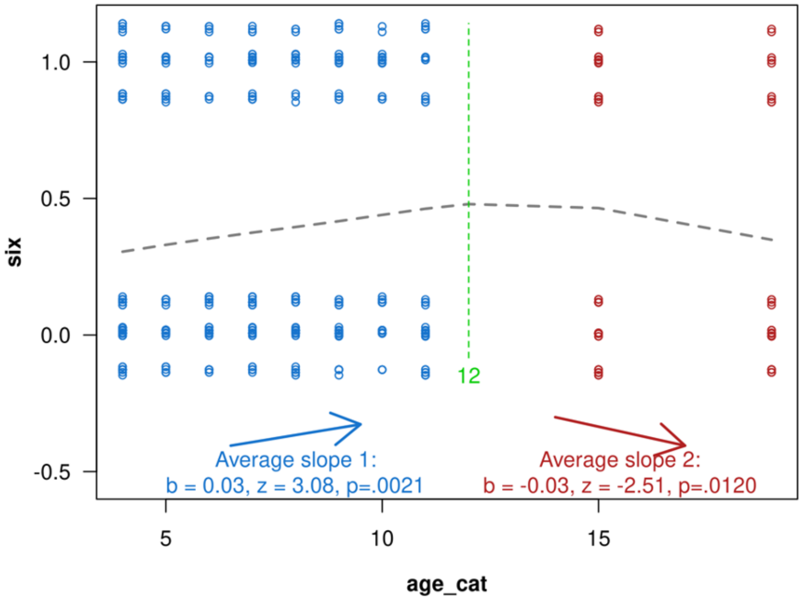
Where n is the number of all possible outcomes, and p_i is the probability of an object being classified to a particular class. In this model, our class of outcomes is binary 0 or 1 for 1 represents cooperation in GB and deception in RT, while 0 represents non-cooperation in GB and Truth in RT. Our predictors are a set of variables listed in Table 1, which include 1) Set of acoustic Features; 2) Set of variables related to text sentiment and emotion from voice; 3) Set of variables related individual game. Figure B1 shows an example of decision tree model in RT Female dataset using acoustic features only. The first node shows the information of the sample and represents the feature that distinguishes deception (1) and truth (0). In this model, we have 66 observations, of which 35 is truthful and 31 is deceptive. The dominate class of this sample is 0. The first node split the sample by pitch with a threshold value of 183.62. The left branch is a subgroup with a pitch smaller than or equal to this threshold, while the right branch represents a subgroup with a pitch higher than the threshold. Among the lower pitch group, there is 22 of truth and 10 of deception. The dominant class in this group is truthful. The next voice feature that distinguishes truthful and deceptive behavior is jitter. The model separates this group into two, one with lower jitter, the other with higher jitter. The model will continue to split each group until it can classify each subgroup with only deceptive or truthful behavior.

Random Forest

To obtain a more robust prediction for our analysis, we use Random Forest as our primary machine learning model (see Luan et al. (2020)). Although decision tree is a good classifier algorithm but it is unstable when predicting future outcomes (see Li and Belford (2002)), as decision tree is prone to over-fitting due to the variance and bias trade-off generated by the maximum depth of the nodes. Alternatively, to limiting the depth of the tree, which reduces variance and increases bias, we can combine many decision trees into a single ensemble model known as the random forest. The random forest is a model made up of many decision trees but rather than just simply averaging the prediction of trees (or so called "Bagging Decision Trees"), this model uses two key concepts to create an ensemble model. First is the random sampling of training data points when building trees, and second is the random subsets of features considered when splitting nodes. In our Random Forest model, we use the same specification as our Decision Tree model explained above.

Appendix C (Chapter 3)

Figure C1: Two-lines Test for Inverted U-shaped Age Effects



Notes: Probability of reporting a six across ages. Ages are censored at 4 and 19.

Simonsohn (2018) proposed the two lines test to detect U-shapes without a functional form assumption. The basic idea is to estimate two regression lines, one for low values and one for high values of the independent variable, and then verify that the two slopes have an opposite sign and are individually statistically significant. The breakpoint between the lines is discovered algorithmically.

We use Simonsohn's app (Simonsohn, 2017), which finds an inverted U-shaped effect of age variable to the probability of rolling a six, while controlling other variables such as the stakes, beneficiaries, female, and parents. The average slope of the effect of age below the age of 12 is 0.03 (p=0.0021) and the average slope of the effect of age above the age of 12 is -0.03 (p=0.012). Please refer to Figure C1 for more details.

References

Abeler, Johannes, Anke Becker, and Armin Falk, “Representative evidence on lying costs,” *Journal of Public Economics*, 2014, *113*, 96–104.

_ , Daniele Nosenzo, and Collin Raymond, “Preferences for Truth-Telling,” *Econometrica*, 2019, *87* (4), 1115–1153.

Adam, Hajo and Adam D. Galinsky, “Enclothed Cognition,” *Journal of Experimental Social Psychology*, 2012, *48* (4), 918–925.

Afifi, Walid A, “Nonverbal Communication.,” 2007.

Akerlof, George A. and Rachel E. Kranton, “Identity and the Economics of Organizations,” *Journal of Economic Perspectives*, March 2005, *19* (1), 9–32.

Allport, Gordon W and Hadley Cantril, “Judging personality from voice,” *The Journal of Social Psychology*, 1934, *5* (1), 37–55.

Andreoni, James and Justin M Rao, “The power of asking: How communication affects selfishness, empathy, and altruism,” *Journal of public economics*, 2011, *95* (7-8), 513–520.

Aribandi, Anurag, Divyanshu Agrawal, and Dipanjan Chakraborty, “Note: Evaluating Trust in the Context of Conversational Information Systems for new users of the Internet,” in “ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies (COMPASS)” 2022, pp. 574–578.

Ayal, Shahar, J/’er/’emy Celse, and Guy Hochman, “Crafting messages to fight dishonesty: A field investigation of the effects of social norms and watching eye cues on fare evasion,” *Organizational Behavior and Human Decision Processes*, 2019, *In Press.*

Basilevsky, Alexander T, *Statistical factor analysis and related methods: theory and applications*, Vol. 418, John Wiley & Sons, 2009.

Battigalli, Pierpaolo and Martin Dufwenberg, "Guilt in games," *American Economic Review*, 2007, *97* (2), 170–176.

Belk, Russell W, "Halloween: An Evolving American Consumption Ritual," *ACR North American Advances*, 1990.

Bénabou, Roland and Jean Tirole, "Incentives and prosocial behavior," *American Economic Review*, 2006, *96* (5), 1652–1678.

_ and _ , "Identity, morals, and taboos: Beliefs as assets," *The Quarterly Journal of Economics*, 2011, *126* (2), 805–855.

Benjamini, Yoav, Abba M Krieger, and Daniel Yekutieli, "Adaptive linear step-up procedures that control the false discovery rate," *Biometrika*, 2006, *93* (3), 491–507.

Berns, Gregory S., Emily Bell, C. Monica Capra, Michael J. Prietula, Sara Moore, Brittany Anderson, Jeremy Ginges, and Atran Scott, "The price of your soul: neural evidence for the non-utilitarian representation of sacred values," *Philosophical transactions of the Royal Society of London. Series B, Biological sciences.*, 2012, *367* (1589), 754–762.

Bicchieri, Cristina, Eugen Dimant, and Silvia Sonderegger, "It's Not A Lie if You Believe the Norm Does Not Apply: Conditional Norm-Following with Strategic Beliefs," May 2020. Working Paper.

_ , _ , Simon Gächter, and Danielle Nosenzo, "Observability, Social Proximity, and the Erosion of Norm Compliance," March 2020. CESifo Working Paper, No. 8212, Center for Economic Studies and ifo Institute (CESifo), Munich.

Biener, Christian and Aline Waeber, "Would I Lie to You? How Interaction with Chatbots Induces Dishonesty," *Behavioral & Experimental Economics eJournal*, 2021.

Bodner, Ronit and Drazen Prelec, "Self-signaling and diagnostic utility in everyday decision making," *The psychology of economic decisions*, 2003, *1* (105), 26.

Bohnet, Iris and Bruno S Frey, "Social distance and other-regarding behavior in dictator games: Comment," *American Economic Review*, 1999, *89* (1), 335–339.

Brandtzaeg, Petter Bae and Asbjørn Følstad, "Why people use chatbots," in "Internet Science: 4th International Conference, INSCI 2017, Thessaloniki, Greece, November 22-24, 2017, Proceedings 4" Springer 2017, pp. 377–392.

Brocas, Isabelle and Juan D. Carrillo, "Self-serving, altruistic and spiteful lying in the schoolyard," *Working paper*, 2019.

Bucciol, Alessandro and Marco Piovesan, "Luck or cheating? A field experiment on honesty with children," *Journal of Economic Psychology*, 2011, *32* (1), 73–78.

_ **and** _ , "Luck or cheating? A field experiment on honesty with children," *Journal of Economic Psychology*, 2011, *32*, 73–78.

Burke, Peter, *The Historical Anthropology of Early Modern Italy: Essays on Perception and Communication*, Cambridge University Press, 2005.

Cadsby, C Bram, Ninghua Du, and Fei Song, "In-group Favoritism and Moral Decision-making," *Journal of Economic Behavior & Organization*, 2016, *128*, 59–71.

Cappelen, Alexander W, James Konow, Erik Ø Sørensen, and Bertil Tungodden, "Just luck: An experimental study of risk-taking and fairness," *American Economic Review*, 2013, *103* (4), 1398–1413.

Capra, C Mónica, "Understanding decision processes in guessing games: a protocol analysis approach," *Journal of the Economic Science Association*, 2019, *5* (1), 123–135.

Capra, C. Mónica, Bing Jiang, and Yuxin Su, "Altruistic self-concept mediates the effects of personality traits on volunteering: Evidence from an online experiment," *Journal of Behavioral and Experimental Economics*, 2021, *92*, 101697.

Capra, Monica, Matthew Gomies, and Shanshan Zhang, "The sound of cooperation and deception in high stake events," *Working paper*, 2023.

Celse, Jérémy and Kirk Chang, "Politicians Lie, So Do I," *Psychological Research*, 2017, pp. 1–15.

Charness, Gary and Martin Dufwenberg, "Promises and partnership," *Econometrica*, 2006, *74* (6), 1579–1601.

Chawla, Nitesh V, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, 2002, *16*, 321–357.

Cingl, Lubomír and Václav Korbel, "External validity of a laboratory measure of cheating: Evidence from Czech juvenile detention centers," *Economics Letters*, 2020, *191*, 109094.

Civile, Ciro and Sukhvinder Obhi, "Students Wearing Police Uniforms Exhibit Biased Attention towards Individuals Wearing Hoodies," *Frontiers in Psychology*, 2017, *8*, 62.

Clot, Sophie, Gilles Grolleau, and Lisette Ibanez, "Smug Alert! Exploring Self-licensing Behavior in a Cheating Game," *Economics Letters*, 2014, *123* (2), 191–194.

Cohn, Alain and Michel André Maréchal, "Laboratory Measure of Cheating Predicts School Misconduct," *The Economic Journal*, 03 2018, *128* (615), 2743–2754.

_ , Ernst Fehr, and Michel André Maréchal, "Business Culture and Dishonesty in the Banking Industry," *Nature*, 2014, *516* (729), 86–89.

_ , Michel André Maréchal, and Thomas Noll, "Bad Boys: How Criminal Identity Salience Affects Rule Violation," *The Review of Economic Studies*, 06 2015, *82* (4), 1289–1308.

_ , Michel André Maréchal, and Thomas Noll, "Bad Boys: How Criminal Identity Salience Affects Rule Violation," *The Review of Economic Studies*, 2015, *82* (4 (293)), 1289–1308.

_ , Tobias Gesche, and Michel André Maréchal, "Honesty in the Digital Age," *Management Science*, 2022, *68* (2), 827–845.

Cordellieri, P, M Boccia, L Piccardi, D Kormakova, LV Stoica, F Ferlazzo, C Guariglia, and AM Giannini, "Gender differences in solving moral dilemmas: emotional engagement, care and utilitarian orientation," *Psychological studies*, 2020, *65* (4), 360–369.

Croson, Rachel, Terry Boles, and J Keith Murnighan, "Cheap talk in bargaining experiments: lying and threats in ultimatum games," *Journal of Economic Behavior & Organization*, 2003, *51* (2), 143–159.

den Assem, Martijn J Van, Dennie Van Dolder, and Richard H Thaler, "Split or steal? Cooperative behavior when the stakes are large," *Management Science*, 2012, *58* (1), 2–20.

DePaulo, Bella M, Deborah A Kashy, Susan E Kirkendol, Melissa M Wyer, and Jennifer A Epstein, "Lying in Everyday Life.," *Journal of Personality and Social Psychology*, 1996, *70* (5), 979.

_ **, James J Lindsay, Brian E Malone, Laura Muhlenbruck, Kelly Charlton, and Harris Cooper**, “Cues to deception.,” *Psychological bulletin*, 2003, *129* (1), 74.

Diekmann, Andreas, Wojtek Przepiorka, and Heiko Rauhut, “Lifting the Veil of Ignorance: An Experiment on the Contagiousness of Norm Violations,” *Rationality and Society*, 2015, *27* (3), 309–333.

Dreber, Anna and Magnus Johannesson, “Gender differences in deception,” *Economics Letters*, 2008, *99* (1), 197–199.

_ **and** _ , “Gender Differences in Deception,” *Economics Letters*, 2008, *99* (1), 197–199.

Ekici, Sami, Turgut Kavas, Yaman Akbulut, and Abdulkadir Sengur, “Deception detection From speech signals,” in “International Conference on Advances and Innovations in Engineering (ICAIE)” 2017.

Erat, Sanjiv and Uri Gneezy, “White lies,” *Management Science*, 2012, *58* (4), 723–733.

_ **and** _ , “White Lies,” *Management Science*, 2012, *58* (4), 723–733.

Evans, David, Michael Kremer, and Muthoni Ngatia, “The Impact of Distributing School Uniforms on Children’s Education in Kenya,” Technical Report, World Bank March 2008.

Falk, Armin, “Facing yourself–a note on self-image,” *Journal of Economic Behavior & Organization*, 2021, *186*, 724–734.

Farrell, Joseph and Matthew Rabin, “Cheap talk,” *Journal of Economic perspectives*, 1996, *10* (3), 103–118.

Feil, DK, “How Venetians Think About Carnival and History,” *The Australian Journal of Anthropology*, 1998, *9* (2), 141–162.

Feine, Jasper, Stefan Morana, and Alexander Maedche, “Designing a Chatbot Social Cue Configuration System,” 2019.

Feldman, Robert S, James A Forrest, and Benjamin R Happ, “Self-presentation and Verbal Deception: Do Self-presenters Lie More?,” *Basic and Applied Social Psychology*, 2002, *24* (2), 163–170.

Fischbacher, Urs and Franziska Föllmi-Heusi, “Lies in Disguise - An Experimental Study on Cheating,” *Journal of the European Economic Association*, 2013, *11* (3), 525–547.

Friesdorf, Rebecca, Paul Conway, and Bertram Gawronski, “Gender Differences in Responses to Moral Dilemmas: A Process Dissociation Analysis,” *Personality and Social Psychology Bulletin*, 2015, *41* (5), 696–713. PMID: 25840987.

G., Caceda R. Capra C. M. Berns G. S. Ekins W., “You cannot gamble on others: dissociable systems for strategic uncertainty and risk in the brain,” *Journal of Economic Behavior and Organization*, 2013, *94*, 222–233.

García, Muñoz, Adrián Gil-Gómez de Liaño, Beatriz, and David Pascual-Ezama, “Gender Differences in Individual Dishonesty Profiles,” *Frontiers in Psychology*, 2021, *12.*

Gerlach, Philipp, Kinneret Teodorescu, and Ralph Hertwig, “The truth about lies: A meta-analysis on dishonest behavior.,” *Psychological bulletin*, 2019, *145* (1), 1.

Gino, Francesca and Lamar Pierce, “Dishonesty in the name of equity,” *Psychological science*, 2009, *20* (9), 1153–1160.

_ and _ , “Lying to Level the Playing Field: Why People May Dishonestly Help or Hurt Others to Create Equity,” *Journal of Business Ethics*, 2010, *95* (1), 89–103.

— , Shahar Ayal, and Dan Ariely, "Contagion and Differentiation in Unethical Behavior: The Effect of One Bad Apple on the Barrel," *Psychological Science*, 2009, *20* (3), 393–398.

— , — , and — , "Self-serving Altruism? The Lure of Unethical Actions that Benefit Others," *Journal of Economic Behavior & Organization*, 2013, *93*, 285–292.

Glätzle-Rützler, Daniela and Philipp Lergetporer, "Lying and age: An experimental study," *Journal of Economic Psychology*, 2015, *46*, 12–25.

Gneezy, Uri, "Deception: The role of consequences," *American Economic Review*, 2005, *95* (1), 384–394.

Griffin, John M, Samuel Kruger, and Prateek Mahajan, "Did FinTech lenders facilitate PPP fraud?," *Journal of Finance, forthcoming*, 2022.

Guerra, Alice, Emanuela Randon, and Antonello E Scorcu, "Gender and deception: Evidence from survey data among adolescent gamblers," *Kyklos*, 2022, *75* (4), 618–645.

Guo, Yiting, Ximing Yin, De Liu, and Sean Xu, ""She is not just a computer": Gender Role of AI Chatbots in Debt Collection," 2020.

Haley, Kevin J and Daniel MT Fessler, "Nobody's watching?: Subtle cues affect generosity in an anonymous economic game," *Evolution and Human behavior*, 2005, *26* (3), 245–256.

Hancock, Jeffrey T, Jennifer Thom-Santelli, and Thompson Ritchie, "Deception and design: The impact of communication technology on lying behavior," in "Proceedings of the SIGCHI conference on Human factors in computing systems" 2004, pp. 129–134.

Hauch, Valerie, Iris Blandón-Gitlin, Jaume Masip, and Siegfried Ludwig Sporer, "Linguistic cues to deception assessed by computer programs: A meta-analysis," in "Proceedings of the workshop on computational approaches to deception detection" 2012, pp. 1–4.

Hazan, Valerie and Rachel Baker, "Does reading clearly produce the same acoustic-phonetic modifications as spontaneous speech in a clear speaking style?," in "DiSS-LPSS Joint Workshop 2010" 2010.

Houser, Daniel, Stefan Vetter, and Joachim Winter, "Fairness and cheating," *European Economic Review*, 2012, *56* (8), 1645–1655.

_ , _ , and _ , "Fairness and cheating," *European Economic Review*, 2012, *56* (8), 1645–1655.

Hughes, Susan M and Bradley C Rhodes, "Making age assessments based on voice: the impact of the reproductive viability of the speaker.," *Journal of Social, Evolutionary, and Cultural Psychology*, 2010, *4* (4), 290.

_ , Marissa A Harrison, and Gordon G Gallup Jr, "Sex-specific body configurations can be estimated from voice samples.," *Journal of Social, Evolutionary, and Cultural Psychology*, 2009, *3* (4), 343.

Jacobsen, Catrine, Toke Reinholt Fosgaard, and David Pascual-Ezama, "Why do we lie? A practical guide to the dishonesty literature," *Journal of Economic Surveys*, 2018, *32* (2), 357–387.

Jacobusse, Gert and Cor Veenman, "On selection bias with imbalanced classes," in "International Conference on Discovery Science" Springer 2016, pp. 325–340.

Jordan, Jennifer, Elizabeth Mullen, and J Keith Murnighan, "Striving for the Moral Self: The Effects of Recalling Past Moral Actions on Future Moral Behavior," *Personality and Social Psychology Bulletin*, 2011, *37* (5), 701–713.

Kaiser, Henry F, "The application of electronic computers to factor analysis," *Educational and psychological measurement*, 1960, *20* (1), 141–151.

Khan, Uzma and Ravi Dhar, "Licensing Effect in Consumer Choice," *Journal of Marketing Research*, 2006, *43* (2), 259–266.

Lane, Harlan and Bernard Tranel, "The Lombard sign and the role of hearing in speech," *Journal of Speech and Hearing Research*, 1971, *14* (4), 677–709.

Larcker, David F and Anastasia A Zakolyukina, "Detecting deceptive discussions in conference calls," *Journal of Accounting Research*, 2012, *50* (2), 495–540.

Lasarov, Wassili and Stefan Hoffmann, "Social Moral Licensing," *Journal of Business Ethics*, 2018, *0* (0), 0.

Lassalle, Amandine, Delia Pigat, Helen O'Reilly, Steve Berggen, Shimrit Fridenson-Hayo, Shahar Tal, Sigrid Elfström, Anna Råde, Ofer Golan, Sven Bölte et al., "The EU-emotion voice database," *Behavior research methods*, 2019, *51* (2), 493–506.

Lavan, Nadine, A Mike Burton, Sophie K Scott, and Carolyn McGettigan, "Flexible voices: Identity perception from variable vocal signals," *Psychonomic bulletin & review*, 2019, *26* (1), 90–102.

Lee, Sangwon, Naeun Lee, and Young June Sah, "Perceiving a Mind in a Chatbot: Effect of Mind Perception and Social Cues on Co-presence, Closeness, and Intention to Use," *International Journal of Human–Computer Interaction*, 2020, *36* (10), 930–940.

Li, Ruey-Hsia and Geneva G Belford, "Instability of decision tree classification algorithms," in "Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining" 2002, pp. 570–575.

Li, Stan Z. and Anil Jain, eds, *Fundamental Frequency, Pitch, F0*, Boston, MA: Springer US,

Linton, Ralph and Ailin Linton, "Halloween through Twenty Centuries," 1950.

Lohse, Tim and Salmai Qari, "Gender differences in face-to-face deceptive behavior," *Journal of Economic Behavior & Organization*, 2021, *187*, 1–15.

Luan, Jing, Chongliang Zhang, Binduo Xu, Ying Xue, and Yiping Ren, "The predictive performances of random forest models with limited sample size and different species traits," *Fisheries Research*, 2020, *227*, 105534.

Lundquist, Tobias, Tore Ellingsen, Erik Gribbe, and Magnus Johannesson, "The aversion to lying," *Journal of Economic Behavior & Organization*, 2009, *70* (1-2), 81–92.

Maggian, Valeria and Marie Claire Villeval, "Social preferences and lying aversion in children," *Experimental Economics*, 2016, *19* (3), 663–685.

Mandal, Fatik Baran, "Nonverbal communication in humans," *Journal of human behavior in the social environment*, 2014, *24* (4), 417–421.

Markowitz, David M, "Revisiting the relationship between deception and design: A replication and extension of Hancock et al.(2004)," *Human Communication Research*, 2022, *48* (1), 158–167.

Mazar, Nina and Chen-Bo Zhong, "Do Green Products Make Us Better People?," *Psychological Science*, 2010, *21* (4), 494–498.

_ **and Dan Ariely**, "Dishonesty in Everyday Life and Its Policy Implications," *Journal of Public Policy & Marketing*, 2006, *25* (1), 117–126.

_ **, On Amir, and Dan Ariely**, "The Dishonesty of Honest People: A Theory of Self-concept Maintenance," *Journal of Marketing Research*, 2008, *45* (6), 633–644.

McAleer, Phil, Alexander Todorov, and Pascal Belin, "How do you say 'Hello'? Personality impressions from brief novel voices," *PloS one*, 2014, *9* (3), e90779.

Mehrabian, Albert and Morton Wiener, "Decoding of inconsistent communications.," *Journal of personality and social psychology*, 1967, *6* (1), 109.

_ **and Susan R Ferris**, "Inference of attitudes from nonverbal communication in two channels.," *Journal of consulting psychology*, 1967, *31* (3), 248.

Michailidou, Georgia and Valentina Rotondi, "I'd Lie for You," *European Economic Review*, 2019, *118*, 181–192.

Miller, Kimberly A., Cynthia R. Jasper, and Donald R. Hill, "Costume and the Perception of Identity and Role," *Perceptual and Motor Skills*, 1991, *72* (3), 807–813.

Mol, Jantsje M, Eline C van der Heijden, and Jan JM Potters, "(Not) alone in the world: Cheating in the presence of a virtual observer," *Experimental Economics*, January 2020.

Murphy, Peter J and Olatunji O Akande, "Cepstrum-based estimation of the harmonics-to-noise ratio for synthesized and human voice signals," in "International conference on nonlinear analyses and algorithms for speech processing" Springer 2005, pp. 150–160.

National Retail Federation, "Halloween Headquarters Survey: Halloween Spending to Reach $9 Billion," 2018. from https://nrf.com/insights/holiday-and-seasonal-trends/halloween.

Nieken, Petra and Simon Dato, "Compensation and Honesty: Gender Differences in Lying," 2016.

Nisan, Mordecai and Gaby Horenczyk, "Moral Balance: The Effect of Prior Behavior on Decision in Moral Conflict," *British Journal of Social Psychology*, 1990, pp. 29–42.

O'Connor, Jillian JM and Pat Barclay, "The influence of voice pitch on perceptions of trustworthiness across social contexts," *Evolution and human behavior*, 2017, *38* (4), 506–512.

Penczynski, Stefan P, "Using machine learning for communication classification," *Experimental Economics*, 2019, *22* (4), 1002–1029.

Pérez-Rosas, Verónica, Mohamed Abouelenien, Rada Mihalcea, and Mihai Burzo, "Deception detection using real-life trial data," in "Proceedings of the 2015 ACM on International Conference on Multimodal Interaction" 2015, pp. 59–66.

Pfattheicher, Stefan, Simon Schindler, and Laila Nockur, "On the impact of Honesty-Humility and a cue of being watched on cheating behavior," *Journal of Economic Psychology*, March 2019, *71*, 159–174.

Phutela, Deepika, "The importance of non-verbal communication," *IUP Journal of Soft Skills*, 2015, *9* (4), 43.

Pisanski, Katarzyna, Paul J Fraccaro, Cara C Tigue, Jillian JM O'Connor, and David R Feinberg, "Return to Oz: Voice pitch facilitates assessments of men's body size.," *Journal of Experimental Psychology: Human Perception and Performance*, 2014, *40* (4), 1316.

Preston, Stephanie D and Frans BM De Waal, "Empathy: Its ultimate and proximate bases," *Behavioral and brain sciences*, 2002, *25* (1), 1–20.

Re, Daniel E, Jillian JM O'Connor, Patrick J Bennett, and David R Feinberg, "Preferences for very low and very high voice pitch in humans," *PloS one*, 2012, *7* (3), e32719.

Sachdeva, Sonya, Rumen Iliev, and Douglas L Medin, "Sinning Saints and Saintly Sinners: The Paradox of Moral Self-regulation," *Psychological Science*, 2009, *20* (4), 523–528.

Sanchez, Jafeth E., Andrew Yoxsimer, and George C. Hill, "Uniforms in the Middle School: Student Opinions, Discipline Data, and School Police Data," *Journal of School Violence*, 2012, *11* (4), 345–356.

Secilmis, Erdem, "An Experimental Analysis of Moral Self-regulation," *Applied Economics Letters*, 2018, *25* (12), 857–861.

Sen, Umut Mehmet, Veronica Perez-Rosas, Berrin Yanikoglu, Mohamed Abouelenien, Mihai Burzo, and Rada Mihalcea, "Multimodal Deception Detection using Real-Life Trial Data," *IEEE Transactions on Affective Computing*, 2020.

Serra-Garcia, Marta, Eric Van Damme, and Jan Potters, "Hiding an inconvenient truth: Lies and vagueness," *Games and Economic Behavior*, 2011, *73* (1), 244–261.

Shalvi, Shaul, "Dishonestly increasing the likelihood of winning," *Judgment and Decision making*, 2012, *7* (3), 292–303.

Simonsohn, Uri, "Two Lines Test App," http://webstimate.org/twolines/ 2017.

_ , "Two lines: A valid alternative to the invalid testing of U-shaped relationships with quadratic regressions," *Advances in Methods and Practices in Psychological Science*, 2018, *1* (4), 538–555.

Sindermann, Cornelia, Peng Sha, Min Zhou, Jennifer Wernicke, Helena S Schmitt, Mei Li, Rayna Sariyska, Maria Stavrou, Benjamin Becker, and Christian Montag, "Assessing the attitude towards artificial intelligence: Introduction of a short measure in German, Chinese, and English Language," *KI-Künstliche intelligenz*, 2021, *35*, 109–118.

Snyder, Hannah, Lars Witell, Anders Gustafsson, and Janet R. McColl-Kennedy, "Consumer lying behavior in service encounters," *Journal of Business Research*, 2022, *141*, 755–769.

Sondhi, Savita, Ritu Vijay, Munna Khan, and Ashok K Salhan, "Voice analysis for detection of deception," in "2016 11th International Conference on Knowledge, Information and Creativity Support Systems (KICSS)" IEEE 2016, pp. 1–6.

Sterba, Richard, "On Hallowe'en," *American Imago*, 1948, *5* (3), 213–224.

Summers, W Van, David B Pisoni, Robert H Bernacki, Robert I Pedlow, and Michael A Stokes, "Effects of noise on speech production: Acoustic and perceptual analyses," *The Journal of the Acoustical Society of America*, 1988, *84* (3), 917–928.

Tabachnick, Barbara G, Linda S Fidell, and Jodie B Ullman, *Using multivariate statistics*, Vol. 5, Pearson Boston, MA, 2007.

Traunmüller, Hartmut and Anders Eriksson, "Acoustic effects of variation in vocal effort by men, women, and children," *The Journal of the Acoustical Society of America*, 2000, *107* (6), 3438–3451.

Turmunkh, Uyanga, Martijn J Van den Assem, and Dennie Van Dolder, "Malleable lies: Communication and cooperation in a high stakes TV game show," *Management Science*, 2019, *65* (10), 4795–4812.

Veblen, Thorstein, "The theory of the leisure class.(republished, 1934, by new york: Modern library)," 1899.

Wakin, Malham M, *Integrity First: Reflections of a Military Philosopher*, Lexington Books, 2000.

Walker, Jonathan, "Gambling and Venetian Noblemen c. 1500 - 1700," *Past & Present*, 02 1999, *162* (1), 28–69.

Ward, Donald J, "Halloween: An Ancient Feast of the Dead that Will Not Die," *Folklore and Mythology*, 1981, *1* (1), 4–6.

Wiltermuth, Scott S, "Cheating More When the Spoils are Split," *Organizational Behavior and Human Decision Processes*, 2011, *115* (2), 157–168.

Winkler, Louis and Carol Winkler, "Thousands of Years of Halloween," *New York Folklore Quarterly*, Sep 01 1970, *26* (3), 204.

Yong, An Gie, Sean Pearce et al., "A beginner's guide to factor analysis: Focusing on exploratory factor analysis," *Tutorials in quantitative methods for psychology*, 2013, *9* (2), 79–94.

Zhang, Shanshan, Matthew Gomies, Narek Bejanyan, Zhou Fang, Jason Justo, Li-Hsin Lin, Rainita Narender, and Joshua Tasoff, "Trick for a treat: The effect of costume, identity, and peers on norm violations," *Journal of Economic Behavior Organization*, 2020, *179*, 460–474.

Zheng, Zhuoyuan, Yunpeng Cai, and Ye Li, "Oversampling method for imbalanced classification," *Computing and Informatics*, 2015, *34* (5), 1017–1037.