Journal of Behavioral and Experimental Economics 112 (2024) 102279

Contents lists available at ScienceDirect

Journal of Behavioral and Experimental Economics

journal homepage: www.elsevier.com/locate/jbee

# Would I lie to you? How interaction with chatbots induces dishonesty ☆

Christian Biener [a,*], Aline Waeber [b]

[a] University of St.Gallen, Tannenstrasse 19, CH-9000, St. Gallen, Switzerland
[b] University of St.Gallen, Switzerland

ARTICLE INFO

*Keywords:*
honesty
chatbot
communication
lying cost
agency

ABSTRACT

Is dishonesty more prevalent in interactions with chatbots compared to humans? Amidst the rise of artificial intelligence, this question holds significant economic implications. We conduct a novel experiment where participants report the outcome of a private, payout-relevant random draw to either a chatbot or a human counterpart, with varying degrees of signaled agency. We find that signaling agency increases honesty when interacting with humans but not with chatbots. Moreover, participants are consistently more honest with humans in the presence of agency cues. Our results suggest that social image concerns and perceived honesty norms play a more prominent role in human interactions. Surprisingly, standard online forms generate the same levels of honesty as human-to-human chat interactions. These findings offer valuable insights for designing effective communication and trust-building mechanisms in digital economies where human-chatbot interactions are increasingly prevalent.

## 1. Introduction

Does human economic behavior change when interacting with a bot as opposed to another human being? This a key behavioral question yet to be understood in the context of the vast substitution of humans with automated and autonomous conversational agents (i.e., chatbots) in the digital economy. While conversational artificial intelligence (AI)[1] has already enabled the use of such chatbots for automation of service and assistance tasks with significant efficiency gains (for both consumers and organizations),[2] most recent developments in this field, such as the release of the large language model ChatGPT, showcase the tremendous potential for future disruption (Heilweil, 2022; Mollick, 2022).

However, there is a potential flip side of replacing human-to-human with human-to-bot interactions: humans might be more inclined to report dishonestly when interacting with a bot as opposed to a human, because lying might simply evoke lower psychological costs of lying in the interaction with a bot as opposed to a human. Such psychological costs of dishonest behavior have been linked to social-norm compliance (Krupka & Weber, 2013) as well as social- (Bursztyn & Jensen, 2017) and self-image (Mazar, Amir, & Ariely, 2008) concerns. A key question in addressing this issue is whether non-human bots can be designed to be perceived as more human in a sense that it increases the psychological costs of dishonest behavior when interacting with them, which appears unlikely given the fact that humans expect other humans (but not bots) to possess a mind. A central dimension in this context has been suggested to be the perceived "agency" (Gray & Wegner, 2012). We interpret "agency" specifically as the perceived capacity of the counterpart to exert self-control and make autonomous decisions. This is distinct from the perceived independence or self-sufficiency of the machine from its human creators or operators. Our focus is on whether the interaction signals the presence of a mind capable of deliberate action and self-regulation, rather than on the organizational or physical distance from human oversight. This interpretation aligns with the

☆ We thank two anonymous reviewers, Faith Neale, Kai Barron, Johannes Binswanger, Christian Alexander Hildebrand, and Matthias Weber as well as the participants of the 11th International Conference of the French Association of Experimental Economics (ASFEE), the 2020 Economic Science Association (ESA) Global Meeting, the Annual Meeting of the German Association for Experimental Economic Research 2021, the 2022 Annual Meeting of the American Risk and Insurance Association (ARIA), the Ph.D. in Finance (PiF) Seminar, the Graduate Programme in Economics and Finance (GPEF) poster session 2022, and the 9th Research Seminar of the Institute of Insurance Economics at the University of St.Gallen for their helpful comments and discussions.

* Corresponding author.
*E-mail addresses:* christian.biener@unisg.ch (C. Biener), aline.waeber@unisg.ch (A. Waeber).

[1] In fact, a study by Juniper (2019) estimates that the operational costs savings from using chatbots will reach about USD 8 billion in banking and amount to USD 1.3 billion in insurance claims management alone in 2023.

[2] An insurance example is the US digital insurance company *Lemonade*, which sells insurance policies and settles claims via chatbots or the Swiss insurance company *Helvetia*, which utilizes chatbots in the claims settling process. Several studies expect this technology to entirely change the means of interactions in both banking (Brewster, 2016) and insurance (Hall, 2017).

https://doi.org/10.1016/j.socec.2024.102279
Received 12 January 2024; Received in revised form 24 May 2024; Accepted 6 August 2024
Available online 10 August 2024

psychological dimensions of agency discussed by Gray, Gray, and Wegner (2007) but is operationalized here in the context of self-control during interactive communication.[3]

Considering the economic costs caused by dishonest behavior, reducing dishonesty effectively is of great relevance to firms, policy makers, and the economy in general alike (Mazar & Ariely, 2006). Dishonesty poses a severe negative externality to markets, generating high costs of doing business, which can ultimately lead to the unraveling of markets (Waeber, 2021). A prominent example of everyday dishonesty is insurance fraud, which is estimated to be exceeding USD 40 billion per year, for example, in the US non-health insurance market alone, increasing premiums for an average family to between USD 400 and USD 700 annually (FBI, 2020). Claims handling processes are the key focus of insurers' operational efficiency programs across the value chain (Adler, Shapiro, Finkelstein, & Sterner, 2019) and chatbots are a vital tool for claims process automation already applied by incumbent firms such as Helvetia Insurance (with chatbot Clara) and insurtechs such as Lemonade Insurance Company (with chatbot Jim).

We assess dishonesty in an incentivized behavioral experiment consisting of a one-shot individual reporting of an unobserved random draw via a written online chat interaction following a strict protocol and a two-by-two factorial design with randomized assignment, varying the nature of interaction (i.e., human versus bot) and the degree of mind attribution through a signal of agency. To eliminate the confounding effects from differences in presentation and framing, we originate a homogeneous online chat interface that varies only regarding the type of interaction—i.e., either an automated chatbot or another human. To avoid participants' apprehension about experimenter manipulation, we rely on an external non-physical and verifiable source of uncertainty—decimals of stock index values, which can be easily looked up via a Google widget on the internet. This source of uncertainty at the same time offers the advantages of featuring an objective (uniform) probability distribution that allows us to infer honest reporting at the treatment population level and is available for all web-enabled devices, making it usable for online experimentation. Precluding the concern (and heterogeneity herein) of facing experimenter judgment of reporting behavior as could incur in a standard physical laboratory setting (e.g., at the payout stage), we ensure the anonymity and non-observability of participants when conducting the experiments not in the laboratory, but at home in an online setting. This latter design feature mimics real-world settings of chatbot interactions and thus magnifies external validity.

We observe a material and significant amount of dishonest reporting behavior in our data, unconditional on treatment assignment. The average reported outcome is significantly higher than the expected average outcome under full honesty. Also, the distribution of reports is highly skewed towards the high payoff outcomes, which is again in contrast to the uniform distribution across all possible values expected under full honesty. Our exogenous treatments generate variation in reporting behavior that provides mixed support for our hypotheses. While we find that reporting to a chatbot induces the lowest levels of honesty, we find no statistically different reporting behavior compared to a human counterpart. In contrast, while endowing the human chat interaction with the indication of human agency (i.e., by signaling mind via a variation in the interaction content) causes a significant increase in honesty, endowing the chatbot with an indication of agency does not induce more honest behavior. So, signaling agency in a human counterpart increases honest behavior, while doing so in a chatbot does not. In essence, our results suggest that under online chat-based interactions only the presence of humans that are at the same time able to indicate agency significantly improves honesty outcomes. We are careful, however, with the interpretation of these finding as there is a broad spectrum of alternative ways to potentially vary agency.

Our analysis of pathways further supports the expected role of social- and self-image concerns, and social norms. More precisely, we find that individuals pay less attention to their social image when they interact with a bot principal (even when this bot principal signals agency) as opposed to a human. Furthermore, an experiment in which participants' beliefs about others' behavior were elicited shows that, aside from the higher prevalence of social-image concerns under then human condition, differences in perceived norms might play a role in explaining behavior. Finally, our results suggest that a passive interaction via a standard online form as opposed to a conversational chat interaction could achieve the same levels of honesty as the human-to-human chat interaction. Considering that such online forms are significantly less costly to operate, this is an important finding to consider when designing customer interactions.

The remainder of this paper proceeds as follows. In Section 2, we provide a broad overview of the related literature and derive our hypotheses. Section 3 provides an introduction to the experimental setting, including information on treatment variation and procedures. In Section 4, we present the results, and in Section 5 we conclude.

## 2. Literature and hypotheses

Our work first and foremost contributes to a new field of empirical research focused on understanding the differences between human-to-human and human-to-bot interactions. Nested in this new stream of literature is the fundamental question of the impact of human-to-bot interactions on honest behavior in market transactions. Standard economic theory suggests that individuals would always lie if benefits (i.e., monetary gains) are sufficiently high to cover the costs of dishonest behavior (i.e., punishment). However, the empirical literature documents broad evidence for individual behavior foregoing financial benefits at the benefit of acting honestly in such situations (Abeler, Becker, & Falk, 2014; Cappelen, Sørensen, & Tungodden, 2013; Erat & Gneezy, 2012; Fischbacher & Föllmi-Heusi, 2013; Gneezy, 2005; Gneezy, Kajackaite, & Sobel, 2018; Mazar et al., 2008). In the context of human-to-bot interactions, there is some empirical evidence that some differences might arise in honest behavior, conditional on whether a human interacts with another human or with a chatbot. Hoegen, Stratou, Luces, and Gratch (2015), for example, compare the behavior of people interacting with other humans in face-to-face interactions with their behavior when interacting with virtual humans. They find that subjects are more willing to steal when interacting with virtual humans. Recently, Maréchal, Cohn, and Gesche (2021) show that lying is more prevalent when subjects interact with a machine in the form of a voice response system with pre-recorded voice messages as opposed to a call with another human. In this paper, we focus on a recent key technological innovation – artificial conversational agents or chatbots – now broadly applied by corporations around the globe. Specifically, the current paper extends the recent work of Maréchal et al. (2021) in two ways. First, it studies reactions to chatbots in contrast to a voice response system with pre-recorded voice messages. Second, it elaborates on the process underlying dishonesty with a focus on the role of agency. We state our first hypothesis as follows.

**Hypothesis 1** (*Human Versus Chatbot*)**.** We expect participants to be less honest in an interaction with a chatbot as opposed to an interaction with another human.

[3] The concept of "agency" as used in this study follows the definition provided by Gray et al. (2007), which focuses on the perceived capacity of an entity to exert self-control, intentionality, and planning. This psychological perspective of agency involves attributing mental states and cognitive abilities to the entity. In contrast, the concept of machines being perceived as independent from humans or organizations pertains to the operational autonomy of the machine, i.e., the extent to which the machine operates without direct human intervention or oversight. While agency in our study concerns the internal capabilities and perceived intentionality of the chatbot, perceived independence would focus on external factors such as the machine's operational self-sufficiency and distance from human control.

Despite recent advancements in chatbot technology, no artificial intelligence has yet been developed that has the basic features of what psychologists call a "theory of mind" (Astington, 1994; Baron-Cohen, 2000). A theory of mind is defined as the human ability to attribute mental states such as beliefs, desires, goals, and intentions to others as well as the ability to deceive others (Rabinowitz et al., 2018). This ability enables humans to conceive of the different mental states that other people demonstrate. Such an ability is thus essential for human cognition and social interactions. For example, theory of mind enables us to understand the meaning behind an ironic remark. Evidence on the theory of mind concept mainly originates from the field of cognitive neuroscience. In an early contribution to neuroeconomics, McCabe, Houser, Ryan, Smith, and Trouard (2001) compare neural activity in a trust game with both human and computer counterparts. They find significant increases in activation in medial prefrontal brain regions among cooperators in human-to-human interactions compared to human-to-bot interactions. This brain region is thus involved in integrating theory-of-mind processing with cooperative behavior. A study by Gogoll and Uhl (2018) reveals that humans are reluctant to delegate decisions to a machine when it comes to moral issues. This is supported by studies showing that people are averse to machines making morally-relevant driving, legal, medical, and military decisions, and that this aversion is mediated by the perception that machines lack a full human mind (Bigman & Gray, 2018). Kirchkamp and Strobel (2019) have extended the idea of responsibility sharing to human–machine interactions. In a dictator game, the number of selfish choices increases when the decision is shared with a computer. However, humans behave even more selfishly when the decision is shared with another human. This is mainly due to the fact that dictators perceive themselves as more responsible for the receiver's payoff in the computer setting.

It should be noted that several earlier studies reveal mind attribution of humans when interacting with machines. Nass and Moon (2000) show that it is possible for humans to ascribe human-like attributes to machines. They argue that such behavior is due to "mindless overusage" of human social categories and "overlearned" social behaviors. In a laboratory experiment, Reeves and Nass (1996) find that humans are as polite to computers as they are to humans. Similarly, Fogg and Nass (1997) reveal that individuals apply gender stereotypes to computers even though they are aware of the machine narrator. Furthermore, Tanibe, Hashimoto, and Karasawa (2017) show that imagining a situation in which people treat a robot in a good way made them more willing to attribute a mind to it, and increased mind attribution led to more positive attitudes towards the robot. We state our second hypothesis as follows.

**Hypothesis 2** (*Signaling of Agency*)**.** We expect the endowment of the human counterpart with an indication of agency (i.e., by signaling mind via a variation in the interaction content) to increase honest behavior, while an indication of agency of the chatbot counterpart is not expected to cause this effect, because humans expect other humans (but not chatbots) to possess a mind.

We explore a range of further pathways that could determine honest behavior in our setting. The act of lying violates an individual's self-concept as an honest person (Mazar et al., 2008; Shalvi, Eldar, & Bereby-Meyer, 2012, and recent evidence suggests that individuals also experience social-image concerns as a result of their dishonesty (Abeler, Nosenzo, & Raymond, 2019; Khalmetski & Sliwka, 2019). Recent theoretical models combine both a desire to appear honest with a desire to be honest in the utility function (Abeler et al., 2019; Khalmetski & Sliwka, 2019). People thus not only consider the expected monetary gains from lying, the probability of being caught, and the potential punishment, but also care about not being perceived as a liar by themselves (Waeber, 2021) but also by the receiver of the message at the time of communication. Following this literature, we expect that some participants incur social- or self-image costs of lying as they experience a direct disutility from falsely reporting private information. In the interaction with a bot principal, such social- or self-image costs of lying play a less important role. We thus expect the share of participants who lie maximally (i.e., report high outcomes) to be higher when reporting to a chatbot. Similarly, we expect the share of partial liars (i.e., reporting more credible outcomes) to be higher when subjects report to a human principal. Previous evidence points in that direction. In fact, Maréchal et al. (2021) find that subjects felt more comfortable reporting suspicious success rates (i.e., full lies) when they interacted with a bot compared to a human. Literature generally agrees that when image concerns are weak, liars report the highest feasible number (Khalmetski & Sliwka, 2019). Thus, we argue that the one way to explain the observed differences in partial and full lies are self- or social image concerns, which play a different role depending on the type of interaction partner. We state our third hypothesis as follows.

**Hypothesis 3** (*Social-Image Concerns*)**.** We expect participants to care more about their social image in an interaction with another human as opposed to an interaction with a chatbot, thus reporting less "suspiciously high" values (i.e., seven, eight, or nine) and more "credible" values (i.e., five or six).

Our work relates to additional studies that emphasize important differences in how humans treat machines. De Melo and Gratch (2015) find that recipients in a dictator game expect more money from a machine than from a human, and proposers in an ultimatum game offer more money to a human than to a machine. They show that subjects are more likely to experience guilt in the interaction with a human as opposed to a machine. Gneezy, Imas, and Madarász (2014) demonstrate that individuals who have previously engaged in immoral behavior are more likely to engage in charitable giving. We hypothesize that this *conscience accounting* phenomenon may manifest differently in interactions with humans versus machines when dishonesty is involved. We state our fourth hypothesis as follows.

**Hypothesis 4** (*Guilt*)**.** We expect participants to experience a higher level of guilt in an interaction with another human as opposed to a chatbot and thus we expect higher donations of experimental earnings among those participants in the chatbot interactions as opposed to those in the human interactions.

Previous literature finds that individuals often do not take full advantage of a lying opportunity (Fischbacher & Föllmi-Heusi, 2013; Mazar et al., 2008). This can be explained by the presence of an internalized honesty norm that creates psychological costs of lying (Cohn, Fehr, & Maréchal, 2014; Fischbacher & Föllmi-Heusi, 2013; Krupka & Weber, 2013). We expect that the presence of a bot as opposed to a human interaction changes the perception of the norm (i.e., beliefs of others' behavior), explaining differences in dishonest behavior. We state our fifth hypothesis as follows.

**Hypothesis 5** (*Social Norms*)**.** We expect participants to hold a different social norm regarding honest reporting in that they expect others to be less honest in an interaction with a chatbot as opposed to an interaction with another human.

Although conversational interfaces, such as chatbots, have only recently emerged in the digital realm, traditional online forms have long been a staple for organizations to interact with customers. These forms serve as a relevant control group, representing the established technology widely used in various contexts, and enabling a comparison between their "formal" structure and the more "informal" nature of chatbot interactions. In the absence of an established theoretical framework, we refrain from proposing a specific hypothesis.

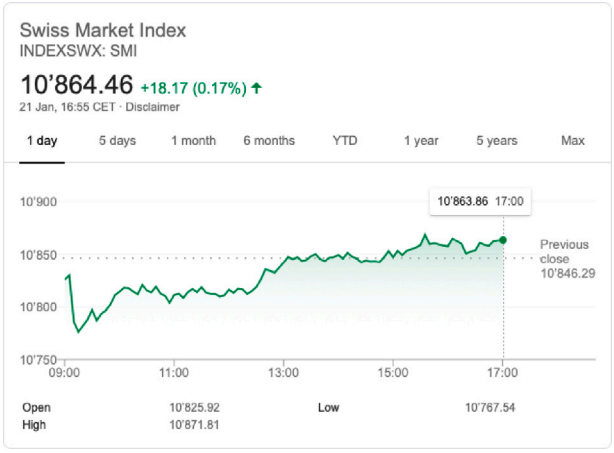


**Fig. 1.** Google widget for SMI.

## 3. Experimental design

Our design follows the fundamental ideas of standard experimental setups to infer dishonest behavior (e.g., Fischbacher and Föllmi-Heusi (2013)) by asking participants to report a randomly generated number, which is unobservable by the experimenter and which determines the payoff for participating in a survey. The context of the survey was framed in the health and risk domain, including questions about health status and health behavior as well as questions about the propensity to take risks in different areas.

In contrast to the use of physical sources of uncertainty such as a die (Fischbacher & Föllmi-Heusi, 2013) or a coin (Maréchal et al., 2021), which can be problematic for online experimentation, we use the value of a stock market index to generate random numbers (see, e.g., (Waeber, 2021)). In particular, the random number is generated by the second decimal place of either the Swiss Market Index (SMI) or the DAX Performance Index (DAX) at a particular point in time—the reported outcome equals the payoff in Swiss Francs (CHF). Participants in the experiment are asked to look up the value of their respective stock index of choice in a Google widget, showing either the SMI or the DAX (see Fig. 1 for an example of the SMI). We then ask them to report the observed random draw via a written chat communication on a computer screen. We emphasize that we will not learn the participants' choice of index, lending further credibility to the unobservability of the source of uncertainty, which is important to avoid reputation and strategic concerns.[4] Because the payoff participants receive for participation depends on the reported outcome, there is a clear incentive to report higher outcomes.

**Table 1**
Treatment conditions.

| | | Principal type | |
|---|---|---|---|
| | | Human | Bot |
| Agency | No | $C_{HU}$ | $C_{BO}$ |
| | Yes | $C_{HU:AGENCY}$ | $C_{BO:AGENCY}$ |

By design of the experiment, we cannot detect dishonesty at the individual level, but, because we know the empirically realized distribution of values,[5] we can infer dishonesty for different subpopulations along our dimension of treatment variation.

We follow Gneezy et al. (2014) and further administer an incentivized measure of guilt to assess whether and how significantly the feeling of guilt plays a differential role between human and bot interactions. Here, we ask participants how much of their proceeds from the experiment they would like to donate to UNICEF (i.e., an amount in the range [CHF 0, CHF 1] in ten centime increments). We use this range so that the donation amount is relatively small compared to the amount that can be gained through dishonesty. We further elicit the subjects' risk attitudes using the survey questions developed by Dohmen et al. (2011).[6]

### 3.1. Treatment variations

We implement four treatments in a between-subject design as shown in Table 1. Table 2 further shows selected communication elements to illustrate the similarities and differences in the treatment conditions. In order to isolate the impact of the type of principal and the role of agency on dishonest behavior, we follow a standard two-by-two factorial design, varying the dimensions of (1) human versus chatbot and (2) agency versus no agency. While under the human conditions (i.e., $C_{HU}$, $C_{HU:AGENCY}$), participants interact with a student assistant, participants interact with an automated chatbot under the bot conditions (i.e., $C_{BO}$, $C_{BO:AGENCY}$). The chatbot was programmed to be gender and age neutral.

We further vary the dimension of agency (i.e., $C_{HU:AGENCY}$, $C_{BO:AGENCY}$). Under these conditions, we facilitate mind attribution among agents towards the principal. Gray et al. (2007) show that people perceive mind in two dimensions: the perceived capacity to sense and feel (i.e., experience)[7] and the perceived capacity to act and do (i.e., agency).[8] Adult humans are seen as capable of both agency and experience and some claim that agency is the essential feature of human minds. In fact, a bot's ability to think is often quite domain specific. Considering the full spectrum of agency, bots are perceived as having less agency than adult humans (Gray & Wegner, 2012). Endowing computers with experience can lead to a phenomenon that has been described as "the uncanny valley", referring to the increasing discomfort humans feel in encountering robots when they appear to be human-minded. Through an experiment, Gray and Wegner (2012) find

[4] Using an online experiment, we seek to eliminate fears of detection, punishment, or reputational concerns (e.g., when interacting with the lab personnel). To test, whether fear of detection plays a role, we elicit subjects' risk attitudes to test whether there is a relationship between risk aversion and dishonest behavior. If the fear of detection is a relevant issue in our design, we would expect that more risk averse subjects are less dishonest. We further include an interaction term between risk aversion and the bot dummy to see whether subjects are more worried about getting caught in the interaction with a human principal. As shown in Table A.1 of the Appendix A, subjects' risk aversion does not significantly decrease reported outcomes. We additionally find that dishonesty is not significantly different under either condition (i.e., human or bot principal). It can thus be concluded that punishment or reputation concerns do not play a role in our experimental design.

[5] The distribution of second decimals in the range of [0, 9] of the DAX and SMI stock index is not only known approximately but exactly, as we look up the data in 5-min intervals at a later point in time in the respective Google widgets. It is important to emphasize that we cannot observe the same data as the participants do during the time of the experiment. We show in Fig. A.1 of the Appendix A that these values are indeed equally distributed.

[6] Precisely, we asked subjects, "How do you see yourself: Are you generally a person who is fully prepared to take risks or do you avoid taking risks" using an 11-point Likert scale ranging from "not at all willing to take risks" to "very willing to take risks".

[7] Attributed to hunger, fear, pain, pleasure, rage, desire, personality, consciousness, pride, embarrassment, and joy.

[8] Including the capacities of self-control, morality, memory, emotion recognition, planning, communication, and thought.

**Table 2**
Exemplary conversational interactions conditional on treatment.

| | Principal type | |
|---|---|---|
| | Human | Bot |
| Agency No | Hello! I'm a student assistant, helping the experimenters to collect your responses for this survey. Are you ready? | Hello! I'm an automated chatbot, the experimenters programmed to collect your responses for this survey. Are you ready? |
| Agency Yes | 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. | 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. |

that a bot capable of experience causes feelings of unease, even without a humanlike appearance. We thus focus on perceived agency here.

More specifically, in $C_{BO:AGENCY}$, the chatbot was programmed to have a discussion on its own capabilities, showing that it can communicate and think (e.g., that it can process a lot of data and interactions at the same time and that it can mimic conversation-like interactions). Similarly, in $C_{HU:AGENCY}$, the student assistant reflects on his capabilities (e.g., that he/she is providing research assistance). The chat communication software allows for both human and bot interaction through the same chat interface, allowing us to avoid differences in format.[9] We would like to acknowledge a possible limitation in the discriminatory power of our intervention for the case of the chatbot interaction. The second treatment dimension, which we describe as varying "agency", could in theory also be interpreted as signaling the chatbot's experience or sophistication. This aligns with the broader understanding of how participants might perceive the interaction, considering (Gray et al., 2007) definition of agency.

### 3.2. Procedures

We recruited participants from the participant pool of the behavioral lab at the University of St.Gallen. Because some of the treatment conditions of the experiment rely on human interaction, and our design requires that we conduct sessions during the trading hours of the SMI and DAX, participants were asked to select a time slot before taking part in the study. The web-link to the study was sent out in a separate e-mail shortly before the session started. Participants took part in the experiment not in the physical laboratory, but at home in an online setting.[10] We explicitly asked participants to make sure that they were in a quiet place without any distractions when starting the experiment.

During the experiment, we informed participants that the experimental data is anonymized and treated confidentially. Participants first received instructions of the experiment via the experimental software oTree (Chen, Schonger, & Wickens, 2016) and were randomly assigned to one of the four conditions, assuring equal distribution of treatments within one experimental session. We informed participants about the nature of the principal (i.e., chatbot or human), conditional on treatment assignment. The student assistants interacting with participants under the human conditions strictly followed a conversational protocol, ensuring that the content of conversations is comparable between human and bot conditions (see supplementary online material for the complete conversational protocol). In particular, the student assistant copied and pasted the questions and replies from the protocol document one-to-one. Only in very few cases did participants probe the chatbot and deviate from the protocol by sending additional questions or messages. In those cases, the experimenters did not respond to such questions or comments by participants. Fig. 2 shows the chat interface that appeared on the main experiment page in the case of the chatbot interaction treatments. The content shown here is only a fraction of the full interaction, which is documented in the supplementary online material. The average experiment duration (in minutes) is 7.06 for $C_{BO}$, 8.16 for $C_{BO:AGENCY}$, 11.04 for $C_{HU}$ and 13.91 $C_{HU:AGENCY}$. We carefully control for the potential role of these differences in our later analyses and find no meaningful effect. Participants were compensated with a fixed participation fee of CHF 3, plus an additional payout that was conditional on the reported random draw (i.e., ranging between CHF 0 and CHF 9). Payments were sent to the participants' bank accounts the evening of the day of participation. We asked participants for their bank account information on a separate website connected to a separate database that we could not link with the experimental data.

### 3.3. Summary statistics

A total of 381 participants with an average age of 23.5 took part in our experiment. The sample has slightly more male participants (59 percent) as opposed female participants (41 percent). About 56 percent of respondents reported to have between CHF 500 and CHF 1499 at their disposal per month, representing the largest income category in our sample. When asked about their sources of income, 62 percent of respondents indicated their family and 29 percent reported their job as their main source of income. 27 percent of participants had a study major in Management. Table 3 provides summary statistics and balancing tests of our control variables. Our sample appears balanced across treatment conditions as all Kruskal–Wallis equality of means test $p$-values are far from relevant significant thresholds. This is expected due to randomized treatment assignment.

## 4. Main results

Fig. 3 shows the distribution of the reported outcomes in the total sample, unconditional on treatment assignment. We find that the average reported outcome is 6.5 and that the distribution of outcomes is highly negatively skewed (i.e., towards the high payoff outcomes). This suggests that participants systematically report higher outcomes than the ones they had actually seen, which is in contrast to the uniform distribution across all possible values between 0 and 9 that would generate an expected average outcome of 4.5 under full honesty (i.e., a one-sided Kolmogorov–Smirnov test of equality with the uniform distribution gives $p$-value $< 0.001$). Following Fischbacher and Föllmi-Heusi (2013) to characterize our sample population, we calculate the share of honest participants (i.e., those reporting an outcome of 0) to be 26 percent and the share of income maximizers (i.e., those reporting an outcome of 9) to be 14 percent.[11]

The results conditional on treatment assignment presented in Fig. 4 and Table 4 offer mixed support for our hypotheses regarding honesty in interactions with humans and chatbots.

[9] Under the human conditions $C_{HU}$ and $C_{HU:AGENCY}$, the conversation is always led by a student assistant interacting with participants.

[10] Previous research finds that findings from traditional, physical laboratory experiments can be replicated in the online laboratory (Horton, Rand, & Zeckhauser, 2011). Despite initial concerns regarding experimental control, these findings are both internally and externally valid.

[11] Fischbacher and Föllmi-Heusi (2013) estimate the percentage of honest people to be as large as 39 percent and the percentage of participants acting as income maximizers to be 22 percent at maximum.

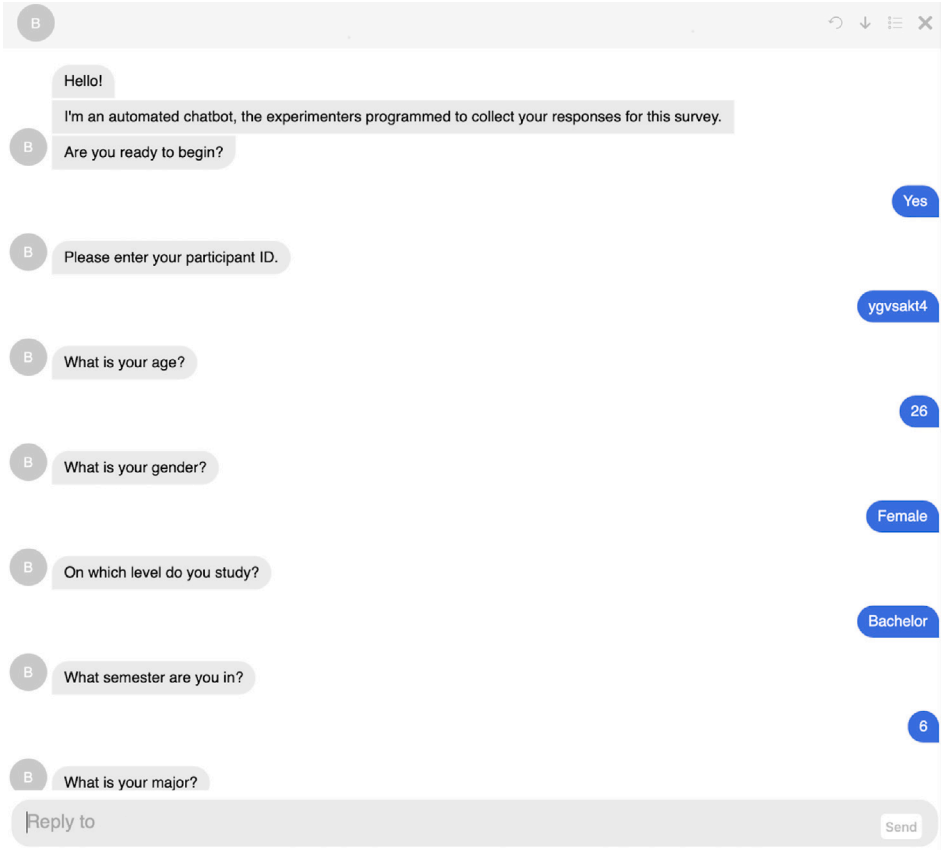


**Fig. 2.** Chat interface.

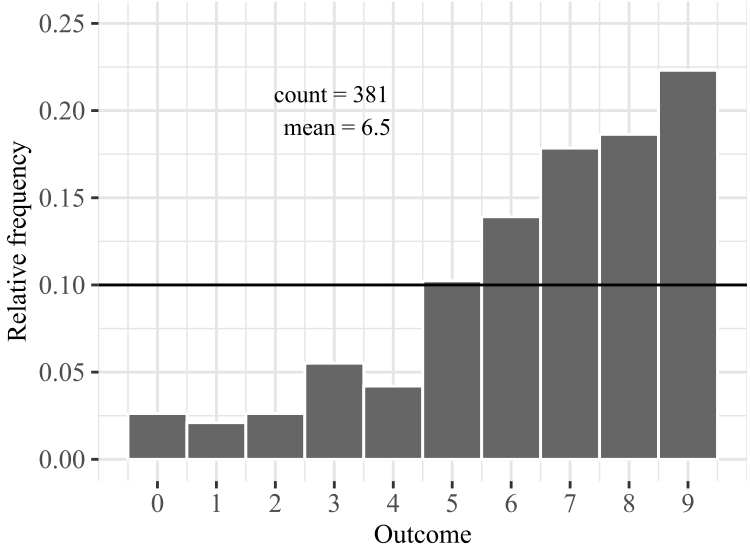


**Fig. 3.** Reports of random draws in total sample. *Note*: The figure shows the distribution of reported random draws unconditional on treatment assignment. The black solid line depicts the uniform distribution expected under truthful reporting.

Our first hypothesis (H1) posited that participants would be less honest in interactions with chatbots compared to humans. The results show no significant difference in reported draws between the $C_{HU}$ and $C_{BO}$ conditions (mean difference = −0.317, $p$-value = 0.494). This suggests that, contrary to our expectation, the mere presence of a human or bot counterpart did not significantly influence honesty in this context.

Our second hypothesis (H2) proposed that signaling agency in a human counterpart would increase honest behavior, while doing so in a chatbot would not have the same effect. The data partially supports this hypothesis. In the human conditions, there was a significant difference between the $C_{HU}$ and $C_{HU:AGENCY}$ conditions (mean difference = 0.674, $p$-value = 0.043). This indicates that signaling agency in a human counterpart did indeed lead to more honest reporting, as we had hypothesized. However, the comparison between the $C_{BO}$ and $C_{BO:AGENCY}$ conditions did not reveal a significant difference (mean difference = 0.147, $p$-value = 0.632). This suggests that signaling agency in a chatbot did not significantly influence honesty, consistent

**Table 3**
Summary statistics and balancing tests.

| | Gender (1) | Age (2) | Income (3) | Income source (4) | Major (5) |
|---|---|---|---|---|---|
| $C_{HU}$ | 0.557 | 23.247 | 0.546 | 0.557 | 0.320 |
| | (0.050) | (0.373) | (0.050) | (0.049) | (0.045) |
| $C_{BO}$ | 0.621 | 23.137 | 0.621 | 0.611 | 0.242 |
| | (0.051) | (0.377) | (0.051) | (0.050) | (0.046) |
| $C_{HU:AGENCY}$ | 0.585 | 24.043 | 0.574 | 0.617 | 0.298 |
| | (0.051) | (0.379) | (0.051) | (0.050) | (0.046) |
| $C_{BO:AGENCY}$ | 0.589 | 23.663 | 0.484 | 0.705 | 0.232 |
| | (0.051) | (0.377) | (0.051) | (0.050) | (0.046) |
| Overall mean | 0.588 | 23.52 | 0.556 | 0.622 | 0.273 |
| *p*-value | 0.844 | 0.167 | 0.288 | 0.203 | 0.453 |
| Observations | 381 | 381 | 381 | 381 | 381 |

*Notes:* The table shows mean values and standard errors of control variables conditional on treatment assignment and overall. We test for equality of means across treatment conditions using the Kruskal–Wallis test; the respective *p*-values of the tests are shown in the second to last line. For ease of interpretation, we recoded our categorial variables into binary variables (i.e., dummy coding) in each case assigning 1 to largest category. Gender is a dummy variable that takes on the value 1 when the participant is male. Income is a dummy variable that takes on the value of 1 if the participant has between CHF 500 to 1499 at his/her disposal per month. Income source is a dummy taking on the value of 1 if their main source of income is their family. Major is a dummy taking on the value of 1 if the participant's major is Management.

**Table 4**
Treatment comparisons.

| Comparison | Mean | *p*-value W | *p*-value KS |
|---|---|---|---|
| $C_{HU} - C_{BO}$ | −0.317 | 0.494 | 0.690 |
| $C_{HU} - C_{HU:AGENCY}$ | 0.674 | 0.043 | 0.028 |
| $C_{HU} - C_{BO:AGENCY}$ | −0.169 | 0.896 | 0.676 |
| $C_{BO} - C_{HU:AGENCY}$ | 0.991 | 0.006 | 0.003 |
| $C_{BO} - C_{BO:AGENCY}$ | 0.147 | 0.632 | 0.740 |
| $C_{BO:AGENCY} - C_{HU:AGENCY}$ | 0.843 | 0.013 | 0.002 |

*Notes*: The table shows mean differences of reported draws defined on the interval between 0 and 9 between all treatment conditions (Column 2) as well as respective *p*-values based on nonparametric Wilcoxon tests (Column 3) as well as two-sample Kolmogorov–Smirnov tests on distributional equality (Column 4).

with our expectation that humans do not attribute agency to chatbots in the same way they do to other humans.

Interestingly, the comparison between $C_{BO:AGENCY}$ and $C_{HU:AGENCY}$ showed a significant difference (mean difference = 0.843, *p*-value = 0.013). This indicates that even when agency is signaled, participants are more honest with humans than with chatbots.

While the results do not fully support our initial hypotheses, they do highlight the complex interplay between the type of counterpart (i.e., human or bot) and the presence of agency cues in shaping honest behavior. The findings suggest that agency cues are more influential when interacting with humans, underscoring the unique social dynamics that arise in human-to-human interactions.

While finding that agency cues are only effective in impacting honest reporting in an interaction with another human, the treatment effect could in principle be driven by mechanisms other than the perception of agency.[12] To test for the validity of attributing the treatment effect to a variation in mind perception, participants were asked to rate the extent to which they perceived their respective counterpart to have agency subsequent to the interaction and to reporting the outcome in the experiment. We measure agency using four survey items.[13] It is important to note that judgment about whether something has a mind or not is often based on perceptions. In the computer science literature, definitions of machine agency often reflect the degree to which machines are perceived to have agency by human actors.

Fig. 5 shows kernel density estimates of perceived agency conditional on treatment assignment. Overall, we find that mean perceived agency is significantly higher under the human conditions as opposed to the bot conditions (*p*-value < 0.001), indicating that humans understand other humans to possess a mind but not bots. As to our intended exogenous variation in agency, we find that indeed it increases perceived agency, but only under the human condition. Mean perceived agency in $C_{HU:AGENCY}$ ($M = 3.314$) is significantly higher (*p*-value = 0.039) compared to $C_{HU}$ ($M = 3.057$). This indicates that our $C_{HU:AGENCY}$ treatment successfully triggered mind perception of participants regarding the other human interaction partner. Under the bot condition, the agency treatment does not impact perceived agency significantly (*p*-value = 0.096). Chatting with a bot thus can – if at all – trigger mind perception only to a limited extent. Related to the possible lack of discriminatory power between agency and other constructs that might be varied by our intervention, this result could suggest that what participants perceive might be more related to the chatbot's experience or level of sophistication rather than its capacity for self-control and planning. Further research could explore these dimensions more specifically to better understand the nuances in human-chatbot interactions.

**Table 5**
Impact of bot nature and agency on reporting high/medium outcomes.

| Comparison | Mean high | *p*-value | Mean medium | *p*-value |
|---|---|---|---|---|
| $C_{HU} - C_{BO}$ | −0.044 | 0.526 | −0.025 | 0.670 |
| $C_{HU} - C_{HU:AGENCY}$ | 0.183 | 0.012 | −0.155 | 0.016 |
| $C_{HU} - C_{BO:AGENCY}$ | −0.055 | 0.431 | −0.004 | 0.944 |
| $C_{BO} - C_{HU:AGENCY}$ | 0.227 | 0.002 | −0.130 | 0.049 |
| $C_{BO} - C_{BO:AGENCY}$ | −0.011 | 0.880 | 0.021 | 0.724 |
| $C_{BO:AGENCY} - C_{HU:AGENCY}$ | 0.238 | 0.001 | −0.151 | 0.020 |

*Notes*: The table shows mean differences of high (i.e., >6) and medium (i.e., 5, 6) reported outcomes between all treatment conditions (Columns 2 and 4) as well as respective *p*-values based on nonparametric Wilcoxon tests (Columns 3 and 5).

### 4.1. Pathways

**Social-image concerns**

In the next step, we aim to examine whether the observed treatment differences in honest behavior can be explained by differences in social-image concerns. The presence of a human principal might cause participants to care more about their social image and thus, they would feel less comfortable telling a lie during the human interaction. To quantify the influence of social-image concerns on dishonesty as well as heterogeneity conditional on treatment assignment, we follow Maréchal et al. (2021) and differentiate the reported outcomes into high (i.e., seven, eight, or nine) – and therefore suspicious – and credible (i.e., five or six).

Table 5 shows treatment effects on reporting high and medium outcomes, respectively, offering mixed support for the hypothesis that participants would be more concerned about their social image when interacting with another human compared to a chatbot. First, the comparison between $C_{HU}$ and $C_{BO}$ reveals no significant difference in the share of individuals reporting high outcomes (−0.044, $p = 0.526$). This suggests that the nature of the interaction partner per se

[12] For example, the specifics of the text elements used to signal agency might have triggered differential reactions not related to agency.

[13] Following the official survey questions on the dimension of mind perception by Gray et al. (2007), participants were asked (1) "To what extent is the counterpart capable of having experiences and being aware of things?", (2) "To what extent is the counterpart capable of understanding how others are feeling?", (3) "To what extent is the counterpart capable of telling right from wrong and trying to do the right thing?", and (4) "To what extent is the counterpart capable of making plans and working towards a goal?" All questions were answered on a five-point Likert scale, ranging from not capable at all (1) to very capable (5). We averaged the final answers to construct an agency index.

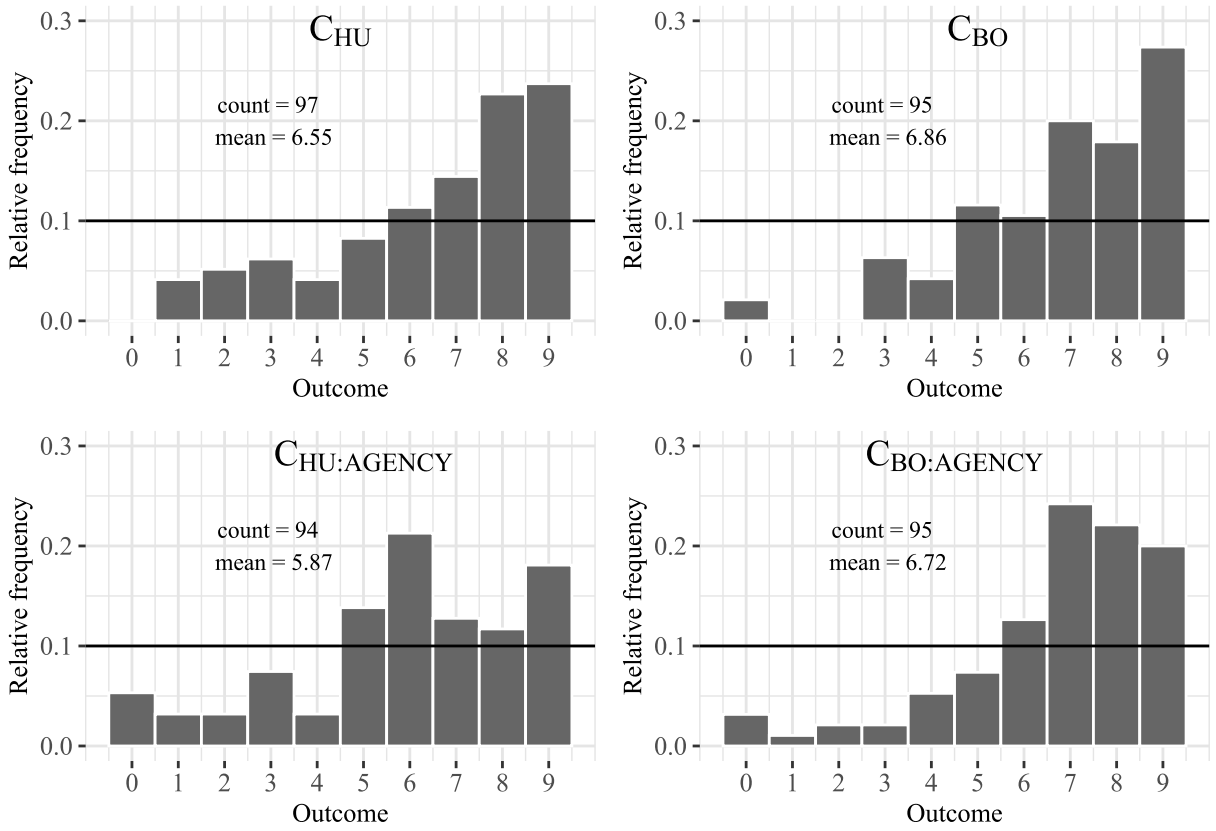


**Fig. 4.** Reports of random draws. *Note*: The figures show the distributions of reported random draws conditional on treatment assignment. The black solid lines depict the uniform distribution expected under truthful reporting.

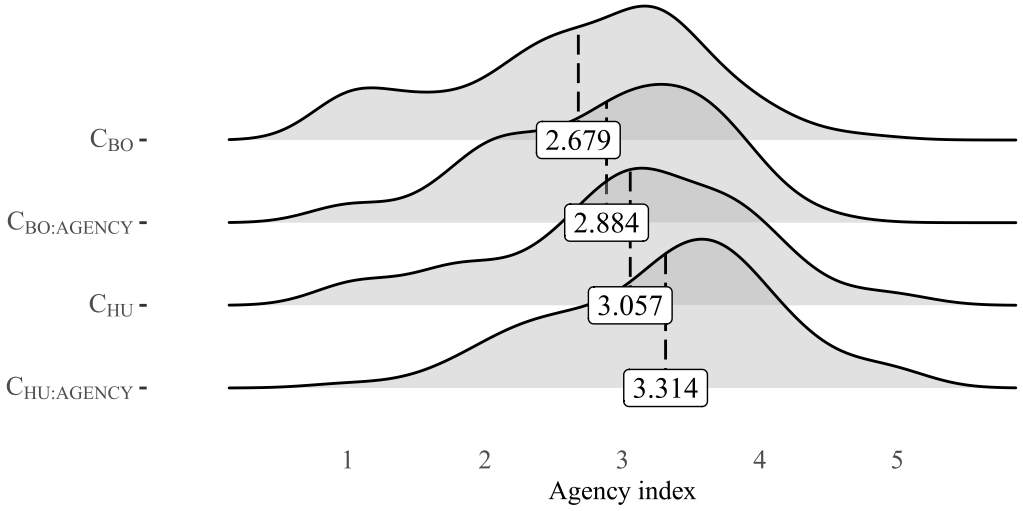


**Fig. 5.** Agency index across conditions. *Note*: The figure shows kernel density estimates of perceived agency conditional on treatment assignment. Dashed lines indicate mean values.

does not influence the likelihood of reporting high outcomes. However, once we endow the human interaction with an indication of agency in $C_{HU:AGENCY}$ we see lower shares of individuals reporting high outcomes both compared to $C_{BO}$ (0.227, $p$-value = 0.002) and $C_{BO:AGENCY}$ (0.238, $p$-value = 0.001). Thus it seems that the indication of agency in the human interaction es key to evoke social-image concerns. This is magnified by the significant direct impact of agency when comparing $C_{HU}$ and $C_{HU:AGENCY}$ (0.183, $p$-value = 0.012). The treatment effects on reporting medium and thus more credible outcomes reveals a substitution of suspicious with credible reportings. In treatments where we see more high outcomes (i.e., human interactions without a signal of agency and bot interactions), we see less medium (i.e., credible) reportings. In particular, $C_{HU}$ has a lower share of individuals reporting medium outcomes compared to $C_{HU:AGENCY}$ (−0.155, $p$-value = 0.016), $C_{BO}$ has a lower share of individuals reporting medium outcomes compared to $C_{HU:AGENCY}$ (−0.130, $p$-value = 0.049), and $C_{BO:AGENCY}$ has a lower share of individuals reporting medium outcomes compared to $C_{HU:AGENCY}$ (−0.151, $p$-value = 0.020).

In summary, our results indicate that individuals pay less attention to their social image when they are interacting with a principal that

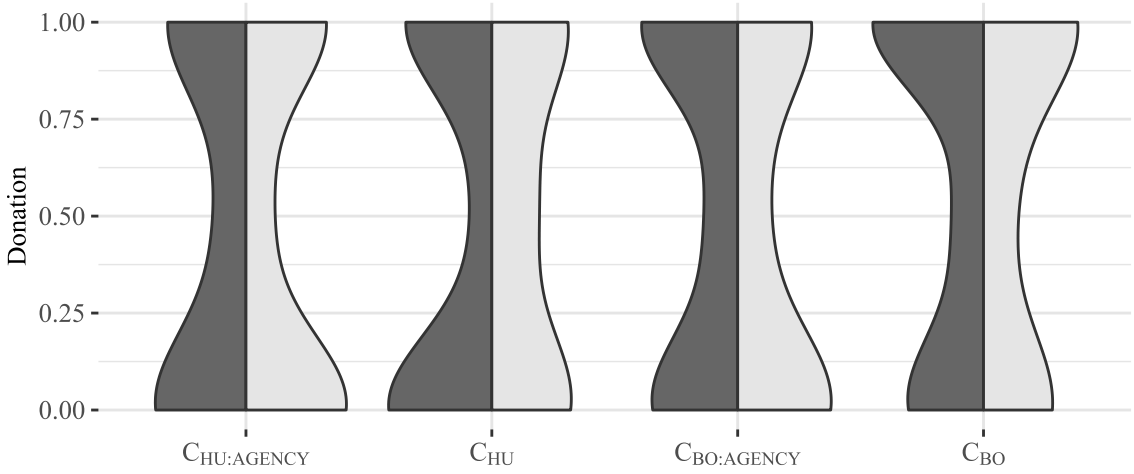


**Fig. 6.** Donations across conditions and reported outcomes. *Note*: The figure shows split violin density plots of donations conditional on treatment assignment and whether participants reported outcomes above 6 (i.e. dark gray) or below or equal to 6 (i.e., light gray).

does not signal agency and that agency does not evoke social-image concerns at all when interacting with a bot principal.

### Guilt

Gneezy et al. (2014) show that prior immoral choices increase the likelihood of charitable giving. We anticipate that such *conscience accounting* might play a differential role between human and machine interactions when it comes to dishonesty. Fig. 6 shows the kernel density estimates of donations to UNICEF (donations could be made in the range [CHF 0, CHF 1] in ten centime increments) conditional on treatment assignment and whether reported outcomes are high (i.e., seven, eight, nine) or low (below or equal to six). Contrary to our expectations and indicated by the symmetry of the split violin density plots, there is no significant variation between donation decisions neither conditional on treatment assignment (i.e., lowest $p$-value of all Kolmogorov–Smirnov tests is 0.165) nor conditional on reported outcomes ($p$-value = 0.79 of correlation between donation and reported outcome); participants who reported higher outcomes are not more likely to donate to charity.

### Social norms

The experimental human and bot conditions generate differences in honest reporting, and here, we examine whether dishonesty becomes more prevalent in the bot interaction due to how bot interactions affect perceptions of norms. In other words, we assume that the presence of a bot interaction partner changes the perception of the norm. We follow the literature on norm elicitation (Cohn et al., 2014; Fischbacher & Föllmi-Heusi, 2013; Kocher, Schudy, & Spantig, 2018; Krupka & Weber, 2013) and conducted an additional experiment with a new set of participants. In the experiment, participants are asked to predict other participants' reporting behavior under both main experimental conditions (i.e., $C_{HU}$ and $C_{BO}$). More specifically, we inform participants that they should predict the behavior of other participants in a previously run experiment (i.e., the "reference experiment"), which we explain in detail. Participants then provide percentage estimates of participants reporting a specific outcome (i.e., from zero to nine), and they are compensated conditional on the accuracy of their predictions. Participants earn CHF 9 if all guesses are correct (i.e., match with the actual behavior observed in the experiments). For every percentage point deviation from the actual shares, payoffs are reduced by CHF 0.1. The minimum payoff in this very short belief elicitation task is CHF 1. Participants additionally receive a show-up fee of CHF 3 that is added to the earnings from the experiment. We recruited participants from the participant pool of the behavioral lab at the University of St.Gallen and excluded all subjects with previous experience in similar experiments. In total, 139 participants took part in this experiment (70 had to predict behavior in $C_{BO}$ and 69 in $C_{HU}$).

Fig. 7 shows the distribution of participants' beliefs about the behavior of others in terms of reporting behavior. On average, we document a prevalent belief that participants in both of our experimental conditions are dishonest at least to some extent ($p$-value $\leq$ 0.001 of Chi-Squared Test). A comparison of the beliefs conditional on treatment assignment (i.e., $C_{HU}$ and $C_{BO}$) shows, however, that the distributions are not significantly different from each other ($p$-value = 0.152 of Kolmogorov–Smirnov test), and participants are expected to be comparably dishonest under both conditions on average. Indeed, the average reported outcome that is expected is virtually identical. An analysis of the dispersion, however, shows that the standard deviation may be higher in $C_{BO}$ ($SD = 5.05$) as opposed to $C_{HU}$ ($SD = 3.83$) implying that the norm may be stronger under $C_{HU}$ (i.e., lower standard deviation of beliefs indicates a higher level of agreement). A Levene's test for homogeneity of variance, however, shows that statistical significance is rather low (i.e., $p$-value = 0.093).

This evidence suggests that the perceived honesty norms do not differ generally between the human and bot conditions. Consequently, as dishonesty is significantly more pronounced under the bot conditions in the actual reference experiment, violations of perceived honesty norms are more prevalent under the bot condition. This is indeed to some extent supported in the data, as the average absolute deviation of actual behavior from the perceived norm is 19.14 percent higher under the bot condition.[14] This suggests that the psychological costs of dishonesty (per unit of deviation from the norm) need to be lower under the bot as opposed to the human conditions.

However, some isolated characteristics of the perceived norms indicate that there might be subtle differences in norms not captured by the above global perspective. First, we find that participants are significantly more likely to anticipate that subjects lie partially under the human condition as opposed to the bot condition by expecting higher shares or reporting the medium outcomes 5 and 6 ($p$-value = 0.013). This is indeed what we have observed in the reference experiment, thus indicating that, aside from the higher prevalence of social-image concerns under the human condition, differences in perceived norms might play a role in explaining behavior. Second, we do not find significantly higher expectations of reporting a maximal outcome of seven, eight, or nine ($p$-value = 0.315) under the bot condition as opposed to the human condition. However, consideration for only the maximal outcome (i.e., nine) indicates marginally significantly

[14] The average absolute deviation is estimated by $\sum_{N=0}^{9} |n_i - a_i|$, whereas $n_i$ represents the expected relative frequency of reporting outcome $i \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$ and $a_i$ the relative frequency of actually reported outcomes $i$ in the reference experiment.

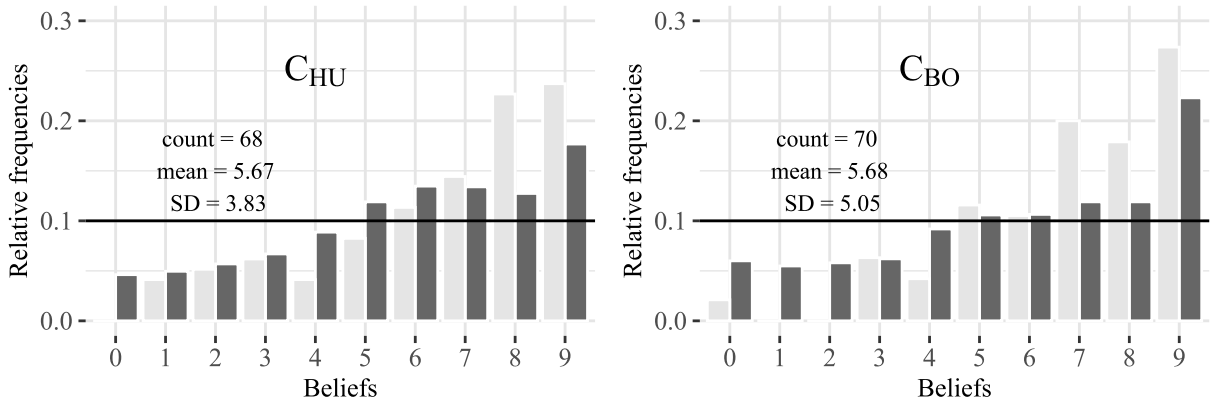


**Fig. 7.** Beliefs about reported outcomes in reference experiment. *Note*: The figures show beliefs about reported payoffs (i.e., the share of participants that is believed to have reported a particular payoff) conditional on treatment in dark gray. The realized relative frequencies in the actual reference experiment are shown in light gray. The black solid lines depict the uniform distribution expected under expectations of truthful reporting.

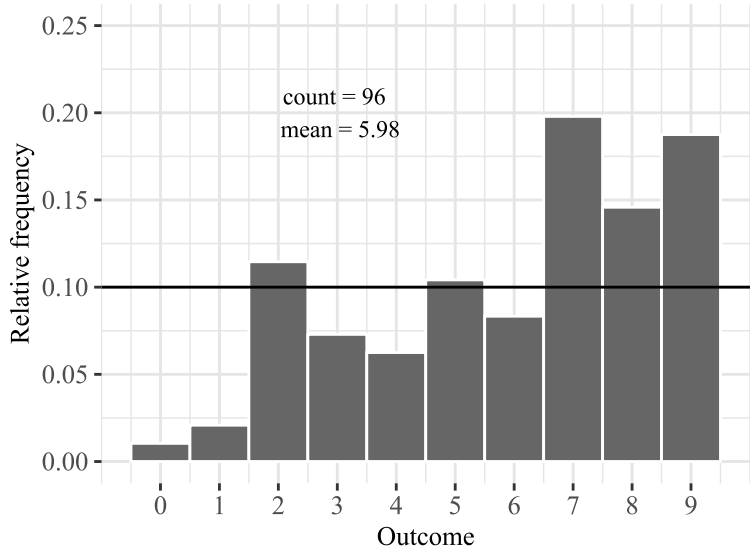


**Fig. 8.** Reports of random draws in online form. *Note*: The figure shows the distribution of reported random draws under the online form treatment condition (i.e., $C_{FO}$). The black solid line depicts the uniform distribution expected under truthful reporting.

higher expectations under the bot condition as opposed to the human condition ($p$-value = 0.071), hinting also in the direction of subtle differences in perceived honesty norms.

**The conversational nature of the interaction**

This research focuses on the impact of the bot nature within a framework of a written conversational interaction (i.e., live chat). While the application of such conversational-like interactions within an online environment is a relatively recent phenomenon, firms have been using online forms for a long time. To evaluate the impact of a written conversational interaction relative to a standard online form on dishonest reporting, we implemented a further treatment ($N$ = 96) in the main experiment, where participants report outcomes in a regular online form without any type of interaction (i.e., $C_{FO}$). All other aspects of the interactions were identical to the treatment conditions, including conversational interactions. Treatment $C_{FO}$ was implemented simultaneously with the other treatments and there were therefore no differences in the subject pool or other timing issues.

Our results (see Fig. 8 and Table 6) suggest that reporting in a standard online form achieves the same levels of honest reporting as the human conversational interaction (i.e., $C_{FO}$ against the intersection of the two sets $C_{HU}$ and $C_{HU:AGENCY}$ gives $p$-value = 0.451). Meanwhile, the online form might even outperform the human chat interaction in the absence of a signal of agency (i.e., $C_{FO}$ against $C_{HU}$ gives $p$-value = 0.101) and is indistinguishable from the human chat interaction with agency (i.e., $C_{FO}$ against $C_{HU:AGENCY}$ gives $p$-value = 0.723). We

**Table 6**
Impact of conversational nature.

| Comparison | Mean | $p$-value | Mean | $p$-value |
|---|---|---|---|---|
| $C_{FO}$ - Human chat | −0.235 | 0.451 | | |
| $C_{FO}$ - Bot chat | −0.810 | 0.012 | | |
| $C_{FO} - C_{HU}$ | | | −0.567 | 0.101 |
| $C_{FO} - C_{HU:AGENCY}$ | | | 0.107 | 0.723 |
| $C_{FO} - C_{BO}$ | | | −0.884 | 0.016 |
| $C_{FO} - C_{BO:AGENCY}$ | | | −0.737 | 0.053 |

*Notes*: The table shows mean differences of reported draws defined on the interval between 0 and 9 between all treatment conditions (Columns 2 and 4) and treatment form as well as respective $p$-values based on nonparametric Wilcoxon tests (Columns 3 and 5).

interpret this finding as an important aspect to be considered when deciding to replace the broadly-applied online forms with written conversational interactions in settings in which dishonesty is detrimental to the firm, such as it is, for example, in the case of insurance claims settlement.

## 5. Conclusion

Digitization is changing the means of interaction of individuals with organizations. In particular, artificial conversational agents (i.e., chatbots) increasingly replace human-to-human interactions with human-to-bot interactions. In this paper, we analyze whether humans are more likely to act dishonestly in an interaction with a chatbot as opposed to another human in a chat communication and which pathways explain differential dishonesty.

We use an innovative experimental design in which participants are asked to report the outcome of an unobserved payout-relevant random draw via a chat interface. While we exogenously vary the type of interaction (i.e., either an automated chatbot or a human chat) and the level of mind attribution (i.e., signal of agency or not), all other aspects of the interaction (e.g., chat interface, content and speed of interaction) were being held constant. The payout-relevant random numbers are generated by a non-physical and verifiable random source of uncertainty—decimals of stock index prices, which can be easily looked up via a Google widget on the internet. We credibly eliminate differences in reputation concerns from the form of interaction conditional on treatment assignment that might otherwise bias potential effects by allowing for the anonymity and non-observability of participants when conducting the experiment not in the laboratory but at home.

Our findings reveal economically relevant differences in honesty depending on whether individuals interact with a chatbot or a human in written chat communication. Signaling agency increases honesty only when interacting with a human, suggesting a differential impact of social cues based on the perceived nature of the counterpart. Notably, the mere transition from human-to-human to human-to-bot interaction does not inherently affect honesty levels. However, we identify a stronger influence of social-image concerns and social norms in human interactions, underscoring the importance of social dynamics in shaping honest behavior. Further analysis reveals that, surprisingly, the conversational style of a written chat with a chatbot, as opposed to a traditional online form, appears to increase dishonest behavior. Given the cost-efficiency of online forms, this finding has important implications for optimizing customer interactions and engagement in digital environments.

The increasing use of chatbots in customer interactions raises important ethical considerations. Our research reveals that while the simple substitution of human interaction with chatbot interaction does not necessarily increase dishonesty, the presence of human agency signals plays a crucial role in promoting honesty. However, similar agency cues from bots may not have the same effect, potentially even leading to increased dishonesty. Surprisingly, chat-style interactions with bots may also contribute to dishonesty compared to traditional online forms. Therefore, organizations should carefully balance the efficiency gains of chatbots with the potential for decreased honesty. This can be achieved through strategic use of chatbots, human-in-the-loop design, transparency, and ongoing research to address emerging ethical challenges in human-bot interactions.

Our conclusions are rooted on the experimental findings presented in this paper, but their interpretation inevitably rests on certain assumptions, because randomizing all potential factors influencing dishonesty under human-to-human and human-to-bot conditions is not possible. For example, we experimentally vary the signaling of agency in a particular way. Future studies could further differentiate the means of mind attribution. Also, we abstract from strong personalizations of the counterpart during the interaction process both for the chatbot and the human. A possible means to affect honest reporting in future studies is to assign social characteristics such as name or gender to the counterpart in the interaction. To our knowledge, there is no extant research showing the relevance of particular (conditional) social characteristics of a chatbot to treat it as if it had its own mind. Finally, the total effect of replacing human-to-human with human-to-bot interactions needs to be evaluated also in the context of the benefits of a human-to-bot interaction that is available 24/7 and allows for a quick and convenient transaction that is valued by consumers. Finally, one might assume that at some point in the future, humans will not be able to distinguish whether they are interacting with real humans or with bots. Future research could therefore examine an *uncertain* setting to see whether humans behave differently when there is uncertainty regarding the type of interaction.

**Table A.1**
Risk aversion and dishonesty.

| | (1) | (2) | (3) |
|---|---|---|---|
| BotTRUE | 0.575** | 0.571** | 0.571** |
| | (0.235) | (0.236) | (0.236) |
| Risk aversion | | −0.047 | |
| | | (0.118) | |
| Risk aversion (Bot = FALSE) | | | −0.067 |
| | | | (0.175) |
| Risk aversion (Bot = TRUE) | | | −0.030 |
| | | | (0.160) |
| Constant | 6.215*** | 6.216*** | 6.217*** |
| | (0.166) | (0.167) | (0.167) |
| Observations | 381 | 381 | 381 |

*Notes:* The table shows (1) OLS estimates of the bot treatment (BO and BO:AGENCY) on the reported random draw defined on the interval between 0 and 9, (2) OLS estimates of the bot dummy as well as a proxy for subjects' risk aversion on the reported random draw, and (3) OLS estimates including an interaction term between risk aversion and the dummy variable stating whether the interaction partner is a human or a bot. *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

## CRediT authorship contribution statement

**Christian Biener:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Aline Waeber:** Writing – review & editing, Writing – original draft, Software, Methodology, Formal analysis, Data curation, Conceptualization.

## Data availability

Data will be made available on request.

## Appendix A

See Table A.1 and Fig. A.1.

## Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.socec.2024.102279.

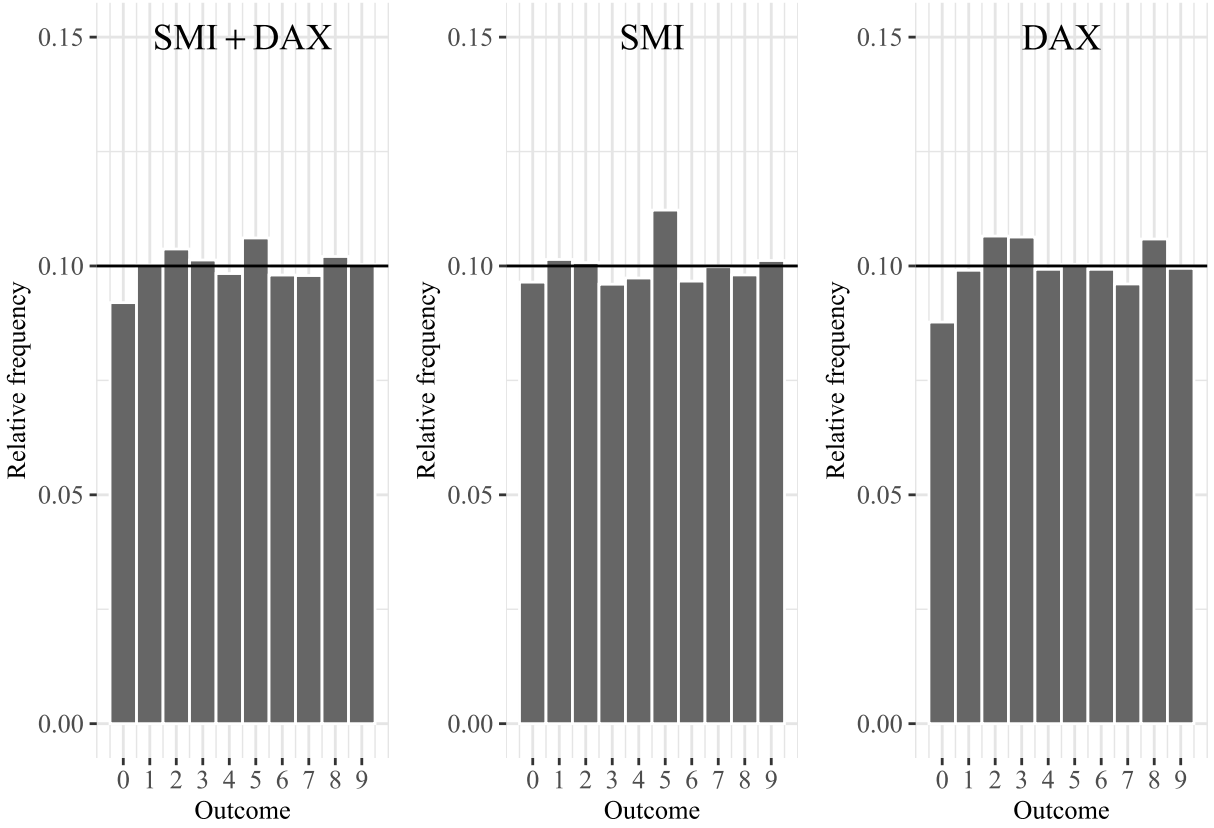


**Fig. A.1.** Distribution of actual random draws from SMI and DAX. *Note*: The figures show actually occurred draws (i.e., the second decimal value) of our two natural randomization devices—the SMI and the DAX stock index value. The data has been collected via daily downloads of HTML files of the respective Google Finance Widgets, containing the actually displayed values of SMI and DAX available to the participants.

## References


Abeler, Johannes, Becker, Anke, & Falk, Armin (2014). Representative evidence on lying costs. *Journal of Public Economics*, *113*, 96–104.

Abeler, Johannes, Nosenzo, Daniele, & Raymond, Collin (2019). Preferences for truth-telling. *Econometrica*, *87*(4), 1115–1153.

Adler, Michael, Shapiro, Scott, Finkelstein, Marc, & Sterner, David (2019). Operational excellence in insurance. https://assets.kpmg/content/dam/kpmg/xx/pdf/2019/03/operational-excellence-report-2019.pdf.

Astington, Janet W. (1994). *The child's discovery of the mind*. Cambridge, MA: Harvard University Press.

Baron-Cohen, Simon (2000). Theory of mind and autism: A review. *International Review of Research in Mental Retardation*, *23*, 169–184.

Bigman, Yochanan E., & Gray, Kurt (2018). People are averse to machines making moral decisions. *Cognition*, *181*(March), 21–34.

Brewster, Signe (2016). Do your banking with a chatbot. *MIT Technology Review*, (May 17), 2016.

Bursztyn, Leonardo, & Jensen, Robert (2017). Social image and economic behavior in the field: Identifying, understanding, and shaping social pressure. *Annual Review of Economics*, *9*, 131–153.

Cappelen, Alexander W., Sørensen, Erik T., & Tungodden, Bertil (2013). When do we lie?. *Journal of Economic Behavior and Organization*, *93*, 258–265.

Chen, Daniel L., Schonger, Martin, & Wickens, Chris (2016). oTree - An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, *9*, 88–97.

Cohn, Alain, Fehr, Ernst, & Maréchal, Michel André (2014). Business culture and dishonesty in the banking industry. *Nature*, *516*(729), 86–89.

De Melo, Celso M., & Gratch, Jonathan (2015). People show envy, not guilt, when making decisions with machines. In *2015 international conference on affective computing and intelligent interaction* (pp. 315–321).

Dohmen, Thomas, Falk, Armin, Huffman, David, Sunde, Uwe, Schupp, Jürgen, & Wagner, Gert G. (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association*, *9*(3), 522–550.

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

FBI (2020). Insurance fraud. https://www.fbi.gov/stats-services/publications/insurance-fraud.

Fischbacher, Urs, & Föllmi-Heusi, Franziska (2013). Lies in disguise-an experimental study on cheating. *Journal of the European Economic Association*, *11*(3), 525–547.

Fogg, B. J., & Nass, Clifford (1997). How users reciprocate to computers. In *Extended abstracts of the CHI97 conference of the ACM/SIGCHI* (pp. 331–332). no. March.

Gneezy, Uri (2005). Deception: the role of consequences. *American Economic Review*, *95*(1), 384–393.

Gneezy, Uri, Imas, Alex, & Madarász, Kristóf (2014). Conscience accounting: Emotion dynamics and social behavior. *Management Science*, *60*(11), 2645–2658.

Gneezy, Uri, Kajackaite, Agne, & Sobel, Joel (2018). Lying aversion and the size of the Lie. *American Economic Review*, *108*(2), 419–453.

Gogoll, Jan, & Uhl, Matthias (2018). Rage against the machine: Automation in the moral domain. *Journal of Behavioral and Experimental Economics*, *74*(March 2017), 97–103.

Gray, Heather M., Gray, Kurt, & Wegner, Daniel M. (2007). Dimensions of mind perception. *Science*, *315*(5812), 619.

Gray, Kurt, & Wegner, Daniel M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition*, *125*(1), 125–130.

Hall, Shanique (2017). How artificial intelligence is changing the insurance industry. In *National association for insurance policy and research - CIPR newsletter 22*. no. August.

Heilweil, Rebecca (2022). AI is finally good at stuff, and that's a problem: Here's why you've been hearing so much about ChatGPT. Vox. https://www.vox.com/recode/2022/12/7/23498694/ai-artificial-intelligence-chat-gpt-openai.

Hoegen, Rens, Stratou, Giota, Luces, Gale M., & Gratch, Jonathan (2015). Comparing behavior towards humans and virtual humans in a social dilemma. In *International conference on intelligent virtual agents* (pp. 452–460).

Horton, John J., Rand, David G., & Zeckhauser, Richard J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental Economics*, *14*(3), 399–425.

Juniper, Rse (2019). Bank cost savings via chatbots. https://www.juniperresearch.com/press/press-releases/bank-cost-savings-via-chatbots-reach-7-3bn-2023.

Khalmetski, Kiryl, & Sliwka, Dirk (2019). Disguising Lies—Image concerns and partial lying in cheating games. *American Economic Journal: Microeconomics*, *11*(4), 79–110.

Kirchkamp, Oliver, & Strobel, Christina (2019). Sharing responsibility with a machine. *Journal of Behavioral and Experimental Economics*, *80*(January), 25–33.

Kocher, Martin G., Schudy, Simeon, & Spantig, Lisa (2018). I Lie? We Lie! why? Experimental evidence on a dishonesty shift in groups. *Management Science*, *64*(9), 3995–4008.

Krupka, Erin L., & Weber, Roberto A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary?. *Journal of the European Economic Association*, *11*(3), 495–524.

Maréchal, Michel André, Cohn, Alain, & Gesche, Tobias (2021). Honesty in the digital age. *Management Science*, *68*(2).

Mazar, Nina, Amir, On, & Ariely, Dan (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, *45*(6), 633–644.

Mazar, Nina, & Ariely, Dan (2006). Dishonesty in everyday life and its policy implications. *Journal of Public Policy & Marketing*, *25*(1), 117–126.

McCabe, K., Houser, D., Ryan, L., Smith, V., & Trouard, T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences*, *98*(20), 11832–11835.

Mollick, Ethan (2022). ChatGPT is a tipping point for AI. *Harvard Business Review.*, https://hbr.org/2022/12/chatgpt-is-a-tipping-point-for-ai.

Nass, Clifford, & Moon, Youngme (2000). Machines and mindlessness: Social responses to computers. *Journal of Social*, *56*(1), 81–103.

Rabinowitz, Neil C., Frank Perbet, H., Song, Francis, Zhang, Chiyuan, Eslami, S. M. Ali, & Botvinick, Matthew (2018). Machine theory of mind. In *Proceedings of the 35th international conference on machine learning* (p. 80). https://arxiv.org/abs/1802.07740.

Reeves, Byron, & Nass, Clifford (1996). Media equation: How people treat computers, television, and new media like real people and places.

Shalvi, Shaul, Eldar, Ori, & Bereby-Meyer, Yoella (2012). Honesty requires time (and lack of justifications). *Psychological Science*, *23*(10), 1264–1270.

Tanibe, Tetsushi, Hashimoto, Takaaki, & Karasawa, Kaori (2017). We perceive a mind in a robot when we help it. *PLoS ONE*, *12*(7), 1–12.

Waeber, Aline (2021). Investigating dishonesty - Does context matter?. *Frontiers in Psychology*, *12*, 1–10.