Knowing More Can Hurt: An Experiment*

Weixuan Zhou†

May 2, 2024

Abstract

Theory suggests that mandatory disclosure of private interest can be harmful, as it deters the transmission of private information. Previous experiments, however, show that disclosing private interest can be beneficial through the psychological effects of *moral licensing* (Cain, Loewenstein and Moore, 2005) and *insinuation anxiety* (Sah, Loewenstein and Cain, 2018). We conducted an experiment to investigate the effect of information disclosure in the setting of strategic information transmission with unknown motives. Our experimental design captures the core spirit of Li and Madarasz (2008), in which a sender has partially aligned interest with a receiver and has private information about his own bias and the state. We show that disclosing private interest results in an essentially unique babbling equilibrium, whereas informative equilibria exist when the private interest is hidden. We then use neologism-proofness (Farrell, 1993) and a best-response dynamics approach to sharpen the theoretical prediction. Our experimental evidence provides support for the theory, as we find that hidden information facilitates information transmission and improves welfare. Meanwhile, our experimental data are inconsistent with the phenomena of *moral licensing* and *insinuation anxiety*. We perform a level-k analysis (Cai and Wang, 2006; Crawford and Iriberri, 2007) as an attempt to explain the payoff differences within and across treatments, and find that a possible source of the welfare loss when the private interest is disclosed is the mismatch between senders and receivers with different levels of sophistication.

Key Words: Information Disclosure, Communication, Information Transmission, Best-Response Dynamics, Laboratory Experiment

*The author is deeply indebted to Wooyoung Lim for his detailed comments, guidance and help. The author thanks Pak Hung Au, Liang Guo, Paul Heidhues, Fuhai Hong, Yucheng Liang, Rui Tang, Qinggong Wu, Xu Zhang, Songfa Zhong, Andreas Ziegler, and participants of the 2023 Stony Brook International Conference in Game Theory, 2023 Econometrics Society European Winter Meeting and 2024 Econometrics Society North American Winter Meeting for their valuable comments and suggestions. The author gratefully acknowledges the financial support from the Department of Economics, Hong Kong University of Science and Technology. The usual disclaimer applies.

†Hong Kong University of Science and Technology, wzhouac@connect.ust.hk

JEL Codes: C91, D82, D83

1 Introduction

In many situations a decision maker lacks decision-relevant information and needs to consult an expert for that information. Oftentimes, the incentives of the expert and the decision maker are not perfectly aligned, and the decision maker does not have a perfect understanding about the expert's motives. For example, a financial professional may provide suggestions to a monetary authority that is contemplating a fiscal policy whose performance will have a direct impact on both the monetary authority and the financial professional. Meanwhile, the professional may have a private interest in an expansionary or contractionary policy, an attitude that may be unknown to the authority. Similarly, a doctor may suggest a surgery to a patient, and the performance of the surgery will affect the well-being of both the doctor and the patient. Meanwhile, the doctor may favor a safe or risky surgery, an attitude that is unknown to the patient. Yet another example comes from the relationship between a sales agent and a customer, in which the agent provides recommendations about products to the customer and may acquire an additional private benefit if the customer purchases certain products.

A question that is of policy interest is whether disclosing the expert's private interest will benefit the expert and the decision maker. It is worthy to note the drastic difference between theoretical predictions and empirical findings on this topic. Theory suggests that mandatory disclosure of private interest can be harmful, as it deters the transmission of private information (Li and Madarasz, 2008). Previous experiments, however, show that disclosing private interest can be beneficial. On one hand, it encourages the expert to send advice even more biased towards his ideal action through the psychological effect of *moral licensing* (Cain, Loewenstein and Moore, 2005). On the other hand, it increases the decision maker's compliance with distrusted advice through the psychological effect of *insinuation anxiety*; that is, the decision maker does not want to reject the expert's preferred proposal for fear that such rejection would be interpreted as a kind of distrust (Sah, Loewenstein and Cain, 2018). Therefore, in the presence of *moral licensing* and *insinuation anxiety*, disclosing the expert's private interest may result in a better outcome at least for the expert.

We conducted an experiment to investigate the effect of disclosing private interest in the setting of strategic information transmission with unknown motives. Our experimental design captures the core spirit of Li and Madarasz (2008), in which a sender's interest is partially aligned with that of a receiver and has private information about his own bias and the state. The sender may prefer an action that is higher or lower than the true state, whereas the receiver prefers an action that is equal to the true

state. We modified the original setting of Li and Madarasz (2008) into a simple, discrete and finite environment to address our research question. In our setting, a sender has private information about the true state, which is randomly drawn from three possible numbers that capture three possible states (high, moderate and low), and his bias, which is either positive or negative. The sender is denoted as a *right sender* if his bias is positive and a *left sender* if his bias is negative. The sender sends a costless and nonverifiable message to the receiver about the state, and the receiver takes an action that affects the payoffs of both parties.

We design four treatments that vary in terms of whether and how the sender's private interest is disclosed. In the first treatment, the private interest is always disclosed. In the second treatment, the private interest is always hidden. In the third treatment, the sender decides whether to disclose his bias to the receiver before observing the state and his bias. In the last treatment, the receiver decides whether to detect the sender's bias before the sender chooses the message. Our treatment variations enable us to examine both the direct effect between information disclosure and nondisclosure and the psychological effects of *moral licensing* and *insinuation anxiety* across different sources of information.

Our theoretical prediction shows that nondisclosure of private interest can facilitate information transmission and benefit both parties. Intuitively, it creates a possibility for a left sender who observes a low state to pool with a right sender who observes a high state, and the remaining left (and also right) senders are pooled together. In this equilibrium, some information is transmitted as the receiver takes an action that is closer to both the the true state and the sender's ideal action when she observes a message from the pooled left (and also right) senders. On the other hand, disclosing private interest deters information transmission, as the sender always has an incentive to exaggerate the message towards his preferred action and the receiver, in turn, downgrades that exaggeration. As a result, essentially no information is transmitted and only the babbling equilibrium (or those essentially equivalent ones) can be realized.

Common to the literature on cheap-talk games, the issue of equilibrium multiplicity arises when the private interest is hidden. To address this issue, we use both neologism-proofness (Farrell, 1993) and a best-response dynamics approach to sharpen the theoretical prediction. The unique equilibrium that survives neologism-proofness is the sender-optimal equilibrium, an outcome that makes both the sender and the receiver better off compared with the babbling equilibrium outcome when the private interest is disclosed. As for the best-response dynamics approach, we adopt a level-k framework in the spirit of

the seminal works such as Cai and Wang (2006) and Crawford and Iriberri (2007). We first specify a level-0 sender to be truthtelling and then iteratively construct the strategies of players with other levels of sophistication. Specifically, for all $k \geq 0$, a level-k receiver best responds to a level-k sender; for all $k \geq 1$, a level-k sender best responds to a level-(k-1) receiver. We find that players' strategies in the level-k model converges to the same sender-optimal equilibrium that survives neologism-proofness as long as $k \geq 1$.

Our experimental evidence provides support for the theory, as we find that hidden information facilitates information transmission and improves welfare. We obtain evidence from both *between-treatment* and *within-treatment* comparisons, and the payoff differences are particularly stark for the data in the last 10 rounds of the experiment. Across all the treatments, both senders and receivers achieve their highest average payoffs in treatment 2, where the private interest is automatically hidden. The differences in average payoffs across treatments are greater and statistically significant in the last 10 rounds of the experiment, for both *between-treatment* and *within-treatment* comparisons. Moreover, within treatment 3, in which senders voluntarily choose whether to reveal their bias, senders who always hide their bias achieve a higher payoff than those who always conceal their bias in the last 10 rounds of the experiment.

Looking at both the aggregate data and the individual level data, we find that our experimental data are inconsistent with the phenomena of *moral licensing* and *insinuation anxiety*. In view that the source of disclosure can also affect players' behavior in a different setting investigated in the literature (Sah, Loewenstein and Cain, 2013), where receivers are less pressured to comply with senders' biased advice when the disclosure of a conflict of interest is revealed by an external source rather than directly from the sender, we also look into whether different sources of disclosure affect players' behavior and payoffs in our setting. Overall, we do not find a clear pattern on how the source of disclosure affects players' strategies.

As an attempt to provide an explanation of the observed payoff differences, we perform a level-k analysis (Cai and Wang, 2006; Crawford and Iriberri, 2007) and characterize players into different levels of sophistication according to their strategies. We show the robustness of our level-k model by comparing players' level-k predicted payoffs with their actual payoffs and find that they are similar. According to our level-k model, we find that a source of welfare loss when the private interest is disclosed is the mismatch of senders and receivers with different levels of sophistication. Specifically, under bias

disclosure, receivers tend to downgrade senders' exaggerated information more than necessary, which results in a welfare loss to both parties and leads to lower average payoffs of both players compared with their nondisclosure counterparts.

The paper is organized as follows. Section 2 reviews related literature. Section 3 provides theoretical background of our experiment. Section 4 presents our equilibrium predictions. Section 5 shows the implementation of our experiment, including design, procedure and hypotheses. Section 6 presents our experimental findings and makes attempts on possible explanations of our findings. Section 7 concludes. Experimental instructions and tables are rendered in the Appendices.

2 Literature Review

Following the seminal work of Crawford and Sobel (1982), there have been studies that theoretically investigate communication games with different focuses. To name a few, Battaglini (2002) examines the setting where the uncertain state of the game has multiple dimensions. Blume, Board and Kawamura (2007) investigate strategic information transmission with communication error. Chen, Kartik and Sobel (2008) propose a NITS (*No Incentive To Deviate*) method to select cheap-talk equilibria. Goltsman, Hörner, Pavlov and Squintani (2009) compare the applicability and performance of mediation, arbitration and negotiation in the setting of strategic information transmission in cheap-talk games. Chakraborty and Harbaugh (2010) examine the setting where an expert uses a multidimensional cheap-talk message to persuade a decision maker. Goltsman and Pavlov (2011) investigate the setting in which a sender talks to multiple receivers. Pei (2015) studies a situation in which a sender gathers costly information before giving advice to a receiver. Sobel (2020) proposes the concepts of lying, deception and damage in strategic settings. In a more recent work, Gordon, Kartik, Lo, Olszewski and Sobel (2023) utilize the concepts of weak dominance and learning to further investigate the issue of equilibrium selection in cheap-talk games.

Among different directions that enrich and generalize the seminal study of Crawford and Sobel (1982), our experiment is closely related to the one on strategic information transmission with unknown motives. To name a few, Ottaviani (2000) compares players' welfare between delegation and communication, Morgan and Stocken (2003) identify the impossibility of a fully revealing equilibrium in a wide class of games, and Dimitrakas and Sarafidis (2005) analyze the size and convergence of equilibria in communication games. Along this stream of literature, our work is most related to the the-

oretical study of Li and Madarasz (2008) on strategic information transmission with unknown motives. Different from their setting where both the message space and the action space are continuous, our experiment adapts the setting into a simple discrete and finite environment. Moreover, when the issue of equilibrium multiplicity is present, we perform equilibrium selections to predict which equilibrium is more likely to happen, which complements the theoretical investigation in Li and Madarasz (2008).

There have also been experimental and empirical studies on strategic information transmission with unknown motives. For example, Cain, Loewenstein and Moore (2005) consider a version of the game where the expert's bias is some positive value whose distribution is unknown to the decision maker, while Koch and Schmidt (2010) consider a version where the expert has some imperfect information about the true state and his payoff function is completely unknown to the decision maker. Both studies find that bias disclosure hurts the well-being of the sender and the receiver; however, neither has a formal game-theoretical model to explain the finding and the settings differ. Sah, Loewenstein and Cain (2013, 2018) find that disclosure of conflict of interests reduces trust but also increases pressure to comply via the panhandler effect and insinuation anxiety. Sachdeva, Iliev and Medin (2009) find that subjects increase their prosocial behavior when they engage in activities that decrease their moral self-concept, and vice versa. Minozzi and Woon (2015) conduct an experiment between two informed experts with opposite biases and an uninformed decision maker. Despite the apparent similarity to our study, there are substantial differences. In our setting, a decision maker is randomly matched with only one expert and receives only one message, whereas in Minozzi and Woon (2015) a decision maker is matched with two experts with opposite motives and receives messages from both of them. Moreover, we adopt a best-response dynamics approach to perform equilibrium selection, which is not utilized in Minozzi and Woon (2015). Perhaps most importantly, in our experiment, we fix the magnitude of the expert's bias to be some constant but create an uncertainty about the direction of the bias, whereas in Minozzi and Woon (2015) the direction of each expert's bias is known but the magnitude of it is uncertain. In this aspect, our work complements the study of Minozzi and Woon (2015).

Our work is also related to experimental investigations in communication games in different settings. To name a few, Blume, DeJong, Kim and Sprinkle (1998) investigate the evolution of the meaning of messages in cheap-talk games and find that meaningful and effective communication endogenously emerges. Weber and Camerer (2003) study the effect of cultural conflict in mergers and find that merging two firms with conflicting cultures results in a decrease in performance. Wang, Spezio and

Camerer (2010) use eyetracking to examine players' behaviors in the lab and find that the eyetracking records of senders are consistent with a level-k model and that players overcommunicate. Battaglini and Makarov (2014) examine the change of behaviors from the addition of a new receiver and find that the observed change of behaviors is consistent with the theoretical predictions. Vespa and Wilson (2016) investigate communications with multiple senders and find that the fully revelation outcomes are realized in certain scenarios while partial revelation outcomes are more common. Hagenbach and Perez-Richet (2018) study a class of sender-receiver disclosure games where senders' incentives may or may not be aligned with the incentives of receivers and find that players' payoffs depend on properties of the incentive graph. Jin, Luca and Martin (2021) use laboratory experiments to test the prediction of disclosure theory. They find that players' actions are strongly related to their beliefs, and senders' beliefs are generally accurate while receivers are insufficiently sceptical about nondisclosed information. Lafky, Lai and Lim (2022) design a series of experiments to systematically investigate the causes of overcommunication and find that the experimental results are in favor of strategic thinking as the primary explanation of overcommunication.

Our work is also related to applications of the best-response dynamics approach in different settings. For instance, Cai and Wang (2006) adopt a level-k approach to identify the presence of overcommunication in a cheap-talk game, Crawford and Iriberri (2007) study the phenomenon of overbidding in an auction model using a level-k model, and Shi and Zillante (2014) study a class of generalized beauty contests using a best-response dynamics approach.

3 Theoretical Background

Our experimental design is motivated by the leading example of Li and Madarasz (2008), which extend the seminal work of Crawford and Sobel (1982) on strategic information transmission to a setting where senders have unknown motives.

A sender is privately informed of the state $\theta \in \Theta = \{1, 3, 5\}$. The common prior is that every state is equally likely. After observing θ, the sender sends a costless and nonverifiable message $m \in M = \{1, 3, 5\}$ about the state to a receiver who then takes an action $y \in Y = \{1, 2, 3, 4, 5\}$. [1]

Assume the receiver's utility function is $U_R(y, \theta) = -(y - \theta)^2$, and the sender's utility function is

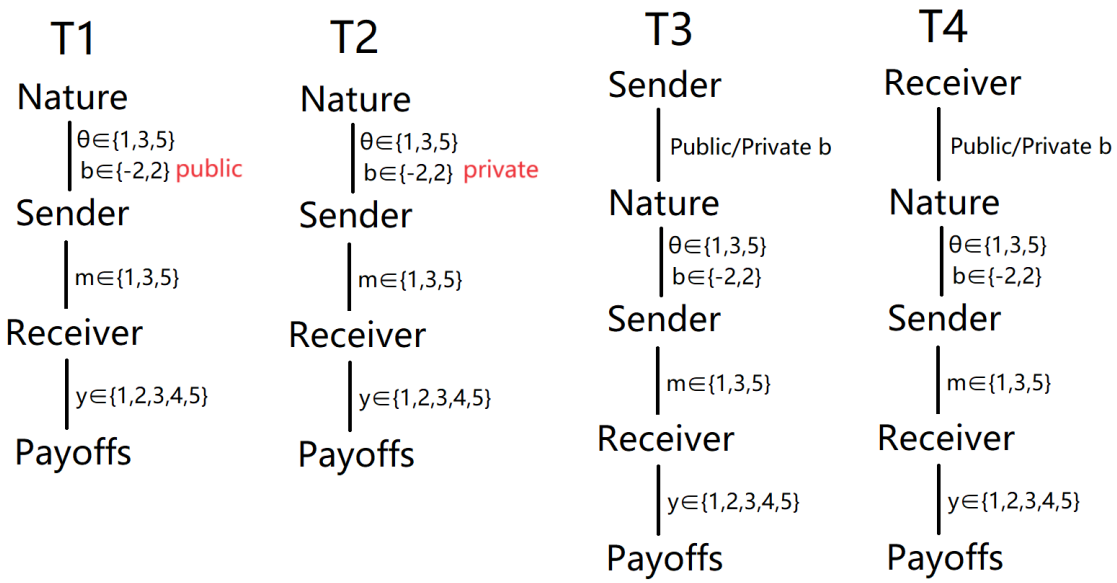
[1] It is without loss of generality to assume the message space has cardinality 3 and the action space has cardinality 5.

$U_S(y,\theta) = -(y-\theta-b)^2$, where $b \in \{-2,2\}$ with equal probabilities. [2] For any $\theta \in \{1,3,5\}$, the receiver's ideal action is $y=\theta$ and the sender's ideal action is $y=\theta+b$. The value of b thus captures the gap between the sender's ideal action and the state, and we shall call this the sender's *bias*. When $b>0$, the sender's ideal action is greater than θ and we say that he is a *right sender*. When $b<0$, the sender's ideal action is less than θ and we say that he is a *left sender*. Both senders and receivers are von-Neumann Morgenstern expected utility maximizers.

Our experiment consists of four treatments: mandatory disclosure, no disclosure, voluntary disclosure and voluntary detection. In the first (mandatory disclosure) treatment, it is common knowledge that both players have perfect information about the sender's bias. In this case, denote the sender's strategy as $\sigma_S(\theta):\Theta\to M$ and the receiver's strategy as $\sigma_D(m):M\to Y$. In the second (no disclosure) treatment, the receiver knows only the distribution of the sender's bias but not the exact value of it. In this case, denote the left sender's strategy as $\sigma_S^L(\theta):\Theta\to M$, the right sender's strategy as $\sigma_S^R(\theta):\Theta\to M$, and the receiver's strategy as $\sigma_D(m):M\to Y$. In the third (voluntary disclosure) treatment, the sender can choose whether to reveal his bias to the receiver. In the last (voluntary detection) treatment, the receiver can choose whether to detect the sender's bias. Both the revelation and detection decisions are made before the bias is realized, which captures the cases where players commit to their revelation and detection decisions independent of the realized value of the bias. Figure 1 demonstrates our design of the four treatments.

[2] Qualitatively, our theoretical results hold when $\frac{3}{2} \le b \le \frac{5}{2}$, in that the sender optimal equilibrium exists, survives neologism-proofness and is the only converging outcome of our best response dynamics approach.

Figure 1: Experimental Treatments



In our experiment, treatment 1 and treatment 2 serve as the benchmark cases for comparison. Treatment 3 consists of two subgames depending on sender's decision on bias revelation: if the sender chooses to disclose his bias, players reach the *disclosure subgame*; otherwise, players reach the *nondisclosure subgame*. Similarly, treatment 4 consists of two subgames depending on receiver's decision on bias detection: if the receiver chooses to detect the sender's bias, players reach the *detection subgame*; otherwise, players reach the *nondetection subgame*. Figures 2 and 3 show the game structures of treatment 3 and 4, respectively.

Figure 2: The Game Structure of Treatment 3

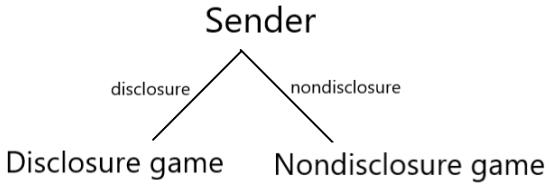
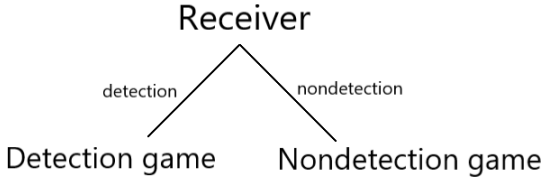


Figure 3: The Game Structure of Treatment 4



In the experiment, we adapt the original setting of Li and Madarasz (2008) to a simple, discrete and finite environment. In particular, we assume that the bias is equally likely to be positive or negative. The distribution assumption accords with the spirit of Li and Madarasz (2008), where the sender's bias takes up to two values. The mean zero property of the distribution captures the case where the sender is neutral on average. For simplicity, we assume that the distribution is symmetric. The solution concept is Perfect Bayesian Equilibrium.

4 Equilibrium Predictions

In this section, we present equilibrium predictions across different treatments and subgames. Section 4.1 presents equilibrium predictions under bias disclosure, which corresponds to treatment 1, the disclosure subgame of treatment 3 and the detection subgame of treatment 4. Section 4.2 presents equilibrium predictions under bias nondisclosure, which corresponds to treatment 2, the nondisclosure subgame of treatment 3 and the nondetection subgame of treatment 4.

4.1 Equilibrium Predictions under Bias Disclosure

Under bias disclosure, by symmetry, it suffices to consider the case in which the receiver interacts with the left sender. An equilibrium is, therefore, characterized by a partition of the state space. There are five possible partitions in total, namely, $\{\{1\},\{3\},\{5\}\}$, $\{\{1,3\},\{5\}\}$, $\{\{1,5\},\{3\}\}$, $\{\{1\},\{3,5\}\}$ and $\{\{1,3,5\}\}$. In each partition, we compute the optimal action of the receiver that corresponds to the message induced by a state (or a state profile), and check whether the sender in each state has any incentive to deviate to a message that corresponds to a different state or deviate to a new message. We find that only the partitions $\{\{1,5\},\{3\}\}$ and $\{\{1,3,5\}\}$ constitute an equilibrium. Note that the receiver will choose $y = 3$ regardless of what message she receives in both partitions, which means that the babbling equilibrium is the essentially unique (in terms of players' expected payoffs) equilibrium outcome under bias disclosure. The sender's expected payoff is -$\frac{20}{3}$, and the receiver's expected payoff is -$\frac{8}{3}$. Proposition 1 summarizes the theoretical prediction under bias disclosure.

Proposition 1 *Under bias disclosure, the babbling equilibrium is the essentially unique equilibrium outcome in terms of players' expected payoffs.*

4.2 Equilibrium Predictions under Bias Nondisclosure

An equilibrium is characterized by a partition of the product space generated from the state space and the space that represents the distribution of the sender's bias, which can be denoted as $T \equiv \{L_1, L_3, L_5, R_1, R_3, R_5\}$. For example, the partition $\{\{L_1\},\{L_3, L_5, R_1\},\{R_3, R_5\}\}$ corresponds to an equilibrium candidate in which the left sender sends a message m_1 when the state is 1 and a message m_2 when the state is 3 and 5, the right sender sends a message m_2 when the state is 1 and a message m_3 when the state is 3 and 5, and the receiver optimally responds by choosing an action 1, 3 or 4

upon receiving m_1, m_2 or m_3, respectively. In the event that the receiver has multiple best responses to any message, we consider all possible cases. It turns out that the game has seven (essentially unique) equilibria, which are summarized in Table 1.

We also compute players' expected payoffs and compare them across different equilibria. Among the 7 equilibria, equilibrium 2 and equilibrium 7 are the Pareto optimal ones. More specifically, equilibrium 2 is sender optimal and equilibrium 7 is receiver optimal.

Proposition 2 summarizes the equilibrium predictions under bias nondisclosure.

Proposition 2 *Under bias nondisclosure, there exist 7 essentially unique equilibria in terms of players' expected payoffs. Among them, there exists a sender-optimal equilibrium and a receiver-optimal equilibrium.*

4.3 Equilibrium Selections under Bias Nondisclosure

To sharpen our theoretical prediction and address the issue of equilibrium multiplicity under bias nondisclosure, we perform equilibrium selections using both neologism-proofness (Farrell, 1993) and a best-response dynamics approach (Cai and Wang, 2006; Crawford and Iriberri, 2007). Both approaches uniquely select the sender-optimal equilibrium.

4.3.1 Neologism-Proofness

In this part, we use the concept of neologism-proofness according to Farrell (1993) for equilibrium selection. For any equilibrium, define $T_S \subset T$ as a self-signaling subset if any sender of type $t \in T_S$ is strictly better off when the receiver acts optimally according to T_S than according to the equilibrium and any sender of a type that does not belong to T_S does not want to induce that action to replace the equilibrium action. An equilibrium is called *neologism-proof* if and only if there does not exist a self-signaling subset.

In Table 2, we construct a self-signaling subset for any equilibrium that is not neologism-proof. As a result, only the sender-optimal equilibrium is neologism-proof. Proposition 3 summarizes our findings.

Proposition 3 *Under bias nondisclosure, only the sender-optimal equilibrium survives neologism-proofness.*

4.3.2 Best-Response Dynamics Analysis

In this part, we perform a best-response dynamics analysis using a level-k model. We assume that each player can be classified into a certain level of sophistication, denoted as a level-k sender or a level-k receiver for some nonnegative integer k. Players' strategies can be iteratively determined once the strategies of level-0 senders are specified. In particular, for any $k \geq 0$, level-k receivers best respond to level-k senders, and level-(k+1) senders best respond to level-k receivers. Crawford, Costa-Gomes and Iriberri (2013) and Blume, Lai and Lim (2017) provide excellent surveys of the applications of level-k analysis in behavioral game theory and in strategic communication games, respectively. Crawford, Costa-Gomes and Iriberri (2013) find that in many communication games, a level-k model with a proper assumption of players' initial behavior nicely characterizes the experimental outcomes. We follow their approach by assuming level-0 senders are truthtelling. We find that players' strategies converge to the sender optimal equilibrium when $k \geq 1$.

Denote the left sender's message when the state is j as L_j, the right sender's message when the state is j as H_j, and the receiver's action when the message j as A_j, where $L_j \in \{1, 3, 5\}, H_j \in \{1, 3, 5\}$ and $A_j \in \{1, 2, 3, 4, 5\}$ for any $j \in \{1, 3, 5\}$. Table 3 presents the best-response dynamics analysis and Proposition 4 summarizes our finding in this part.

Proposition 4 *Under bias nondisclosure, in our level-k model, players' strategies converge to those prescribed by the sender-optimal equilibrium when $k \geq 1$.*

Theorem 1 summarizes our theoretical predictions in this section.

Theorem 1 *Under bias disclosure, the babbling equilibrium is the essentially unique equilibrium. Under bias nondisclosure, there exist multiple equilibria; among them, only the sender-optimal equilibrium is neologism-proof and is the unique converging outcome of our best-response dynamics analysis.*

Note that both senders and receivers are better off in the sender-optimal equilibrium under nondisclosure than in the babbling equilibrium under disclosure. Intuitively, in the sender-optimal equilibrium, a right sender who observes a low state is pooled with a left sender who observes a high state, and the remaining right (also left) senders are pooled together. In this equilibrium, some information is transmitted as the receiver takes an action that is closer to both the the true state and the sender's ideal action when she observes a message from the pooled left (also right) senders. Figures 4 and 5 provide an intuitive illustration of the argument.

Figure 4: Babbling Equilibrium under Disclosure

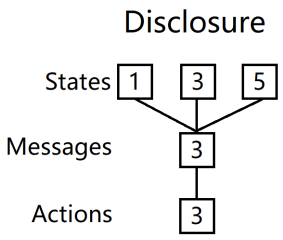
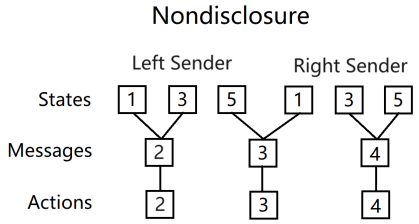


Figure 5: Sender-Optimal Equilibrium under Nondisclosure



5 Experimental Implementation

5.1 Design and Procedure

Our experiment was conducted using oTree (Chen, Schonger and Wickens, 2016) at The Hong Kong University of Science and Technology. A total of 118 undergraduate/postgraduate students with no prior experience of such experiments were recruited as our experimental subjects. Our experiment consisted of four treatments. Each treatment consisted of two identical sessions using a *between-subjects* design. Each subject participated in exactly one session, and each session involved 14 or 16 subjects. All sessions were conducted in November 2022.

Each subject was randomly assigned to be a sender or receiver with equal probability, and the role was fixed throughout the experiment. In each round, a sender was randomly and anonymously matched with a receiver to form a group, and the groups were reshuffled after each round. To begin with, in treatment 3, the sender decided whether to disclose his bias to the receiver; in treatment 4, the receiver decided whether to detect the sender's bias. Then, in all treatments, the sender privately observed the state θ and his bias. After that, in treatment 1, the bias was automatically revealed to the receiver, whereas in treatment 2 the bias was automatically hidden from the receiver. In treatment 3 and 4, the bias was either revealed or hidden depending on the decision of the relevant player. Then, the sender sent a costless and nonverifiable message $m \in \{1, 3, 5\}$ to the receiver. Finally, the receiver took an action $a \in \{1, 2, 3, 4, 5\}$ and each player got his/her payoff. At the end of each round, we provided information feedback on which state was chosen, whether the bias was disclosed (treatment 3) or detected (treatment 4), the sender's bias and his message, the receiver's action and the subject's own payoff.

5.2 Hypotheses

To postulate on players' disclosure (detection) decisions in treatment 3 (4) and on the comparison of subjects' expected payoffs across different treatments and subgames, we calculate players' expected payoffs given their levels of sophistication under bias disclosure and nondisclosure. More specifically, for each $k \geq 0$, a level-(k+1) sender's expected payoff is calculated according to his optimal strategies when he interacts with a level-k receiver, and a level-k receiver's expected payoff is calculated according to her optimal strategies when she interacts with a level-k sender. Finally, we assume a level-0 sender's expected payoff is 0 by assuming that he interacts with a credulous receiver who always takes an action equal to the message. Table 4 and Table 5 present players' actions given their levels of sophistication under bias disclosure. Table 6 and Table 7 summarize the expected payoffs of players with different levels of sophistication.

Our first hypothesis concerns the equilibrium predictions in treatment 1 and 2. We formulate this hypothesis based on Theorem 1 in Section 4.

Hypothesis 1 *In treatment 1, the babbling equilibrium will be realized. In treatment 2, the sender-optimal equilibrium will be realized.*

Our second hypothesis concerns senders' decisions on whether to reveal their bias in treatment 3. We formulate our null hypothesis based on the theoretical predictions and our alternative hypothesis based on a behavioral analysis. Our null hypothesis is that senders will not reveal their bias. First, for any level of sophistication, the sender is weakly better off under bias nondisclosure than under bias disclosure. Second, senders are better off in the sender-optimal equilibrium under bias nondisclosure than in the babbling equilibrium under bias disclosure. Therefore, according to the game structure of treatment 3, senders will not reveal their bias. Our alternative hypothesis is that senders will reveal their bias. This follows from the psychological effects of *moral licensing* and *insinuation anxiety*. Once revealing their bias, senders may choose a message that is more biased towards their ideal action due to the effect of *moral licensing*, and receivers may increase their compliance to the even more biased advice due to the effect of *insinuation anxiety*.

Hypothesis 2 *In treatment 3, senders will not reveal their bias.*

Our third hypothesis concerns receivers' decisions on whether to detect the sender's bias in treatment 4. Our null hypothesis is that receivers will not detect the sender's bias. First, level-2 and above

receivers are strictly better off under bias nondisclosure. Second, receivers are better off in the sender-optimal equilibrium under bias nondisclosure than in the babbling equilibrium under bias disclosure. Therefore, according to the game structure of treatment 4, receivers will not detect the sender's bias. Our alternative hypothesis is that receivers will detect the sender's bias, as level-1 receivers are strictly better off under bias disclosure.

Hypothesis 3 *In treatment 4, receivers will not detect the sender's bias.*

Our fourth hypothesis concerns players' expected payoffs within and across treatments. We postulate that both players achieve a higher expected payoff in treatment 2, the nondisclosure subgame of treatment 3 and the nondetection subgame of treatment 4 than in treatment 1, the disclosure subgame of treatment 3 and the detection subgame of treatment 4, since both senders and receivers are better off in the sender-optimal equilibrium under bias nondisclosure than in the babbling equilibrium under bias disclosure.

Hypothesis 4 *Both players achieve a higher expected payoff in treatment 2, the nondisclosure subgame of treatment 3 and the nondetection subgame of treatment 4 than in treatment 1, the disclosure subgame of treatment 3 and the detection subgame of treatment 4.*

6 Experimental Findings

We present our experimental findings in three parts. In section 6.1, we summarize our findings of subjects' disclosure/detection decisions and compare their average payoffs within and across treatments. In section 6.2, we compare our experimental data with the predictions of the psychological effects of *moral licensing* and *insinuation anxiety* and find that our experimental data are inconsistent with the two effects. We also investigate whether the source of disclosure affects players' behavior but do no find a clear pattern on this. As an attempt to provide an explanation of the observed payoff differences within and across treatments, we perform a level-k analysis in section 6.3. We characterize subjects into different levels of sophistication according to their behavior and calculate their level-k predicted payoffs based on the empirical distribution of subjects' levels of sophistication. Our level-k model suggests that a possible source of welfare loss when the bias is disclosed is the mismatch of senders and receivers with different levels of sophistication.

6.1 Disclosure/Detection Decisions and Payoffs

Table 8 summarizes senders' disclosure decisions in treatment 3 and receivers' detection decisions in treatment 4. It turns out that more than 90% of the receivers choose to detect the sender's bias in treatment 4, an observation that is inconsistent with our null hypothesis in *Hypothesis 3* and in favor of the alternative hypothesis. As for treatment 3, slightly more than half of the senders choose to disclose their bias and the remaining senders choose not to, an observation that neither supports nor rejects our null hypothesis in *Hypothesis 2*.

Table 9 summarizes subjects' average payoffs within and across treatments. Both senders and receivers achieve the highest average payoffs in treatment 2. Meanwhile, within treatment 3, disclosure gives senders and receivers higher average payoffs; Within treatment 4, senders on average earn more with nondisclosure, whereas receivers on average earn more with disclosure.

The payoff differences are more quantitatively and statistically significant in the last 10 rounds of the experiment. Table 10 summarizes subjects' average payoffs across treatments in the last 10 rounds. Players' average payoffs across the four treatments exhibit a clear ranking, which is precisely inverse to the frequency of bias revelation in that treatment. Moreover, the differences are greater in magnitude and are statistically significant at the 0.1 level across different groups of senders (Treatment 1 v.s. treatment 2, p=0.09; treatment 2 v.s. treatment 4, p=0.04; treatment 3 disclosure v.s. treatment 3 nondisclosure, p=0.03, Wilcoxon signed-rank test). As for receivers, the differences are also greater in magnitude and the payoff difference between treatment 1 and treatment 2 is statistically significant at the 0.05 level (Treatment 1 v.s. treatment 2, p=0.03; treatment 2 v.s. treatment 4, p=0.31; treatment 3 disclosure v.s. treatment 3 nondisclosure, p=0.55, Wilcoxon signed-rank test).

Player-level data are also consistent with the findings in the last 10 rounds of the experiment. In treatment 3 in the last 10 rounds of the experiment, 2 senders always disclose their bias and 2 senders never disclose their bias. The average payoffs of the nondisclosing senders are 87.6 and 79.6 whereas the average payoffs of the disclosing senders are 58.0 and 53.2, and the difference is statistically significant at the 0.05 level (p=0.04, Wilcoxon signed-rank test).

6.2 Moral Licensing, Insinuation Anxiety and Source of Disclosure

Psychological studies have shown that disclosing private interests can result in the effects of *moral licensing* and *insinuation anxiety* (Cain, Loewenstein and Moore, 2005; Sah, Loewenstein and Cain,

2018); that is, senders will provide an advice that is even more biased towards their ideal actions and receivers feel more morally obliged to comply with it. Note, however, that these studies are concerned with the setting where senders have a direct conflict of interest with receivers. We would thus like to investigate whether these psychological effects persist in our setting where senders and receivers have partially aligned interests.

In a related study, Sah, Loewenstein and Cain (2013) find that the source of disclosure can also affect players' behavior. In particular, they find that receivers are less pressured to comply with senders' biased advice when the disclosure of a conflict of interest is provided by an external source rather than directly from the sender. In view of this, we also investigate whether the source of disclosure affects players' behavior and payoffs in our setting where players have partially aligned interests. [3]

Tables 11-20 summarize the frequencies of senders' messages and receivers' actions across different treatments/subgames. Tables 21-30 summarize our test results.

6.2.1 Moral Licensing

Our analysis on *moral licensing* consists of both *between-treatment* and *within-treatment* comparisons. Our *between-treatment* comparison involves experimental data from treatment 1 and treatment 2, whereas our *within-treatment* comparison involves experimental data from treatment 3, where senders choose whether to disclose their bias and there are relatively balanced observations of both disclosure and nondisclosure.

Tables 11 and 12 demonstrate whether the effect of *moral licensing* persists from the *between-treatment* comparison. The first two lines of the tables exhibit a similar pattern, while senders choose message 1 more frequently when the bias is negative and the state is 5 in treatment 1. With positive bias, senders choose message 3 more frequently in treatment 2 while choose message 5 more frequently in treatment 1 when the state is 1, and the rest are similar. Our *between-treatment* comparison suggests that senders choose a message that is closer to their ideal action under bias nondisclosure, a finding that is inconsistent with the phenomenon of *moral licensing*.

Tables 13 and 14 demonstrate whether the effect of *moral licensing* persists from the *within-treatment* comparison. With negative bias, when the state is 3, senders choose message 1 more frequently in the nondisclosure subgame while choose message 3 more frequently in the disclosure subgame; when the state is 5, senders choose message 3 more frequently in the nondisclosure subgame while choose message 1

[3] The data analysis of the nondetection subgame is omitted because of limited observations.

more frequenly in the disclosure subgame. With positive bias, senders choose message 3 more frequently in the nondisclosure subgame while choose message 5 more frequently in the disclosure subgame when the state is 1. Overall, our *within-treatment* comparison also suggests that our data are inconsistent with the phenomenon of *moral licensing*.

We also perform statistical testings to investigate the effect of *moral licensing* in our experiment. Table 21 and Table 22 present the Wilcoxon signed-rank test statistics of *between-treatment* and *within-treatment* comparisons of senders' communication strategies, respectively. We compute the differences between senders' messages and their ideal actions for each state, and compare the differences between disclosure and nondisclosure. Table 21 shows that senders choose a message that is *further away* from their ideal actions when the bias is positive and the state is 1 or 3 and when the bias is negative and the state is 5 under disclosure, and there are no statistically significant differences in other cases. Table 22 shows that senders choose a message that is *further away* from their ideal actions when the bias is positive and the state is 1 and when the bias is negative and the state is 3 or 5 under disclosure, and there are no statistically significant differences in other cases. Therefore, Table 21 and Table 22, together, suggest that the senders only exhibit the psychological effect of *moral licensing*, if at all, when their private interest is hidden.

On top of the aggregate data, we also look at the individual level data to see whether the effect of *moral licensing* is present among subjects. In particular, we look at the individual level data in treatment 3, where senders make balanced decisions on whether to reveal their bias. Among the 14 senders, 7 of them choose to reveal their bias in at least 5 rounds and also to hide their bias in at least 5 rounds, and our analysis on *moral licensing* will be concerned with these senders.

The analysis of the individual level data is similar to that of the aggregate data. For each sender, we compute the differences between his message and his ideal action in each round, and then compare the differences under bias disclosure and those under nondisclosure. Similar to our finding in the aggregate data, we find that the individual level data are inconsistent with the effect of *moral licensing*. Among the 7 senders, one sender has systematically smaller differences under disclosure, one sender has systematically smaller differences under nondisclosure, and the other 5 senders do not have a clear pattern on which of the differences are smaller.

Overall, both the aggregate data and the individual level data are inconsistent with the effect of *moral licensing*.

6.2.2 Insinuation Anxiety

Similar to the one on *moral licensing*, our analysis on *insinuation anxiety* consists of both *between-treatment* and *within-treatment* comparisons. Our *between-treatment* comparison involves experimental data from treatment 1 and 2, whereas our *within-treatment* comparison involves experimental data from treatment 3.

Tables 16 and 17 demonstrate whether the effect of *insinuation anxiety* persists from the *between-treatment* comparison. When the message is 3, receivers choose action 3 more frequently in treatment 2 but choose action 5 more frequently in treatment 1, which is to the contrary of the effect of *insinuation anxiety* that receivers choose an action closer to the message when the bias is disclosed. Therefore, our data are inconsistent with the effect of *insinuation anxiety* from the *between-treatment* comparison.

Tables 18 and 19 demonstrate whether the effect of *insinuation anxiety* persists from the *within-treatment* comparison. Again, when the message is 3, receivers choose action 3 more frequently in the nondisclosure subgame but choose action 1 more frequently in the disclosure subgame. On top of that, when the message is 5, receivers choose action 4 more frequently in the nondisclosure subgame but choose action 3 more frequently in the disclosure subgame, which is to the contrary of the effect of *insinuation anxiety* that receivers choose an action closer to the message when the bias is disclosed. Therefore, our data are also inconsistent with the effect of *insinuation anxiety* from the *within-treatment* comparison.

We also perform statistical testings to investigate the effect of *insinuation anxiety* in our experiment. Tables 23 and 24 present the Wilcoxon signed-rank test statistics of *between-treatment* and *within-treatment* comparisons of receivers' communication strategies, respectively. We compute the differences between senders' messages and receivers' actions for each message, and compare the differences between disclosure and nondisclosure. For *between-treatment* comparison, receivers choose an action that is closer to the message when the message is 1 under bias disclosure, while the converse is true and the difference is notably larger in magnitude when the message is 3. For *within-treatment* comparison, receivers choose an action that is closer to the message when the message is 3 under bias nondisclosure. Overall, the aggregate data do not provide systematic support for the presence of the effect of *insinuation anxiety*.

On top of the aggregate data, we also look at the individual level data to see whether the effect of *insinuation anxiety* is present among subjects. In particular, we look at the individual level data in treatment 3, where senders make balanced decisions on whether to reveal their bias. Receivers are

randomly matched with different senders across 20 rounds of the game and thus have balanced chances to observe a revealed or hidden bias. In fact, all the 14 receivers observe a revealed bias in at least 5 rounds and also a hidden bias in at least 5 rounds. Therefore, our analysis on *insinuation anxiety* will be concerned with all the receivers.

The analysis of the individual level data is similar to that of the aggregate data. For each receiver, we compute the differences between her action and the message she receives in each round, and then compare the differences under bias disclosure and those under nondisclosure. Similar to our finding in the aggregate data, we find that the individual level data are inconsistent with the effect of *insinuation anxiety*. For the 14 receivers, 4 of them have systematically *smaller* differences under nondisclosure (which is to the opposite of the phenomenon of *insinuation anxiety*) and 10 of them do not have a clear pattern on which of the differences are smaller.

Overall, both the aggregate data and the individual level data are inconsistent with the effect of *insinuation anxiety*.

6.2.3 Source of Disclosure

Our analysis on the source of disclosure consists of *between-treatment* comparison, which involves experimental data from treatment 1, the disclosure subgame in treatment 3 and the detection subgame in treatment 4.

Tables 11, 13 and 15 demonstrate whether the source of disclosure plays a role in senders' communication strategies. Comparing Table 11 with Table 13, we find that senders choose message 1 more frequently when the bias is negative and the state is 3 in treatment 1, but they choose message 3 more frequently when the bias is positive and the state is 1 or 3 in the disclosure subgame of treatment 3. Comparing Table 11 with Table 15, we find that senders choose message 1 more frequently when the bias is negative and the state is 1 and choose message 3 more frequently when the bias is positive and the state is 1 in the detection subgame of treatment 4. Comparing Table 13 with Table 15, we find that senders choose message 1 more frequently when the bias is negative and the state is 1 or 3, choose message 3 more frequently when the bias is negative and the state is 5, and choose message 3 more frequently when the bias is positive and the state is 1 in the detection subgame of treatment 4. Overall, we find that senders typically choose a message closer to their ideal action when their bias is passively detected compared with the case when their bias is actively disclosed, but there is no clear pattern

when the comparison involves an external source of disclosure.

Tables 16, 18 and 20 demonstrate whether the source of disclosure plays a role in receivers' actions. Comparing Table 16 with Table 18, we find that receivers choose action 1 more frequently when the bias is negative and the message is 1 in treatment 1 but choose action 3 more frequently when the bias is negative and the message is 5 in the disclosure subgame of treatment 3. Table 16 and Table 20 are generally similar, except that receivers choose action 3 slightly more frequently when the bias is negative and the message is 3 in treatment 1. Comparing Table 18 and Table 20, we find that receivers choose action 1 more frequently when the message is 1 regardless of the bias, and choose action 5 more frequently when the bias is negative and the message is 3 or 5 in the detection subgame of treatment 4. Overall, there is no clear pattern on how the source of disclosure affects receivers' strategies.

We also perform statistical testings to investigate the effect of the source of disclosure in our experiment. Tables 25-27 present the Wilcoxon signed-rank test statistics of senders' communication strategies against different sources of disclosure. Overall, we do not find a clear pattern that relates these two. For example, Table 25 shows that senders choose a message closer to their ideal actions when the bias is negative and the state is 3 in treatment 1, while the converse is true when the bias is positive and the state is 1. Table 26 and Table 27 show that senders choose a message that is closer to both the true state and their ideal actions when the bias is negative and the state is 1 in the detection subgame of treatment 4, but the rest are similar.

Tables 28-30 present the Wilcoxon signed-rank test statistics of receivers' actions against different sources of disclosure. Table 28 shows that receivers take an action that is closer to the sender's message when the bias is negative and the message is 1 or 5 in treatment 1 compared with the disclosure subgame of treatment 3, while the converse is true when the bias is negative and the message is 3. Table 29 shows that receivers take an action that is closer to the sender's message when the bias is negative and the message is 3 in treatment 1 compared with the detection subgame of treatment 4, while the converse is true when the bias is positive and the state is 1. Table 30 shows that receivers take an action that is closer to the sender's message when the bias is negative and the message is 1 or 5 and when the bias is positive and the message is 1 in the detection subgame of treatment 4 compared with the disclosure subgame of treatment 3, while the converse is true when the bias is negative and the message is 3. Overall, these results are mixed and do not provide support for the finding in Sah, Loewenstein and Cain (2018) that an external source of disclosure mitigates the effect of *insinuation anxiety*.

Overall, we do not find evidence of systematic patterns on how the source of disclosure affects players' behavior.

6.3 Level-k Analysis

As an attempt to explain the payoff differences within and across treatments, we utilize the level-k model described in Section 5.2 to characterize subjects' observed behavior. A sender is classified as level-0, level-1 or level-2 under bias disclosure and as level-0 or level-1/equilibrium under bias nondisclosure. A receiver is classified as level-0, level-1, level-2 or pooling under bias disclosure and as level-0, level-1/equilibrium or pooling under bias nondisclosure. [4] A subject is classified into a certain level of sophistication if (i) the strategies of the subject are better matched with that level of sophistication than with any other level of sophistication and (ii) the strategies of that level of sophistication match the actual data at least 60% of the times; otherwise, the subject is unclassified. In case there is a tie, a subject is classified into the lowest level of sophistication among them. In treatment 3, a subject is classified separately under bias disclosure and bias nondisclosure, provided that the subject has at least four observations in that category. In treatment 4, a subject is classified based on the observations under bias detection only, since 92.3% of the observations belong to this. Table 31 summarizes our classification method. Based on this method, 75%, 81% and 83% of the subjects in treatment 1, 2 and 4 are classified, and 75% of the subjects in each group in treatment 3 are classified. [5]

Tables 32-36 summarize our level-k classification. Overall, more than 90% of the senders can be classified into at least one category (94.12%), while the fraction of receivers that can be classified into at least one category is 63.01%. This may, in part, be due to the fact that a sender has all the relevant information about the state, his bias and his ideal action to make his decision, whereas a receiver is faced with uncertainty in multiple dimensions. Across all the treatments, most senders are classified as level-1 (70.59%), suggesting that a sender typically chooses a message that is closest to his ideal action. Meanwhile, the classification pattern of receivers varies across different treatments and subgames. Under bias disclosure, most receivers are classified as level-2 (56.00%), while under bias nondisclosure, most are classified as either pooling (52.38%/47.62%) or level-1 (38.10%).

[4] Note that classifying a sender as a babbling type is not helpful to understand his behavior, since a babbling sender matches with any observation with 100% accuracy. Also note that the strategies for both senders and receivers are identical at all levels of sophistication $k \geq 1$ and the equilibrium level under nondisclosure, according to Table 3.

[5] If the tie happens between pooling and level-1 or above, then the receiver is classified as pooling. If the tie happens between pooling and level-0, then we consider both cases and analyze them separately.

Based on the classification results, we calculate players' payoffs when players of different levels of sophistication interact with each other. To do so, we need to specify the off-equilibrium strategies whenever applicable. According to our level-k model, off-equilibrium strategies need to be specified for level-1 and level-2 receivers under bias disclosure. We assume that a level-1 receiver will randomize over all possible actions with equal probability upon receiving the off-equilibrium message 1 (or 5) from a right (or left) sender, and that a level-2 receiver will randomize over all possible actions with equal probability upon receiving the off-equilibrium message 1 or 3 (3 or 5) from a right (left) sender. Our assumptions about the off-equilibrium strategies are consistent with our level-k classification, since we do not impose any restrictions on off-equilibrium strategies when classifying subjects. Table 37 and 38 summarize the results. In each two-dimensional vector, the first entry indicates the sender's payoff and the second entry indicates the receiver's payoff.

As a robustness check, we compare subjects' level-k predicted payoffs with their actual payoffs. To do so, we calculate subjects' level-k predicted payoffs according to the empirical distribution of the subjects' levels of sophistication in each treatment. Table 39 summarizes our results. [6] Overall, the level-k predicted payoffs in Table 39 closely match the actual payoffs in Table 9, suggesting that our level-k classification works reasonably well in explaining players' observed behaviors.

Tables 32-38, together, suggest an explanation of the variation of subjects' payoffs across treatments. In treatment 2, senders are mostly of level-1 and receivers are mostly of level-1 or pooling. When a level-1 sender interacts with a level-1 receiver, their payoffs are $-\frac{10}{3}$ and -2, respectively. When a level-1 sender interacts with a pooling receiver, their payoffs are $-\frac{20}{3}$ and $-\frac{8}{3}$, respectively. In treatment 1, however, senders are mostly of level-1 and receivers are mostly of level-2. Their corresponding payoffs are $-\frac{22}{3}$ and $-\frac{10}{3}$, both of which are lower than their counterparts in treatment 2. Similarly, in treatment 4, senders are mostly of level-1 and receivers are mostly of level-2. Their corresponding payoffs are $-\frac{22}{3}$ and $-\frac{10}{3}$, which are also lower than their counterparts in treatment 2.

Our discussion above implies that, according to our level-k model, a source of welfare loss when a sender's bias is disclosed is the mismatch between senders and receivers with different levels of sophistication. We start with *between-treatment* comparison. Under bias nondisclosure, the majority of interactions come from level-1 senders with level-1 or pooling receivers in treatment 2, resulting in their payoffs to be $(-\frac{10}{3}, -2)$ or $(-\frac{20}{3}, -\frac{8}{3})$. Meanwhile, under bias disclosure, the majority of interactions

[6] Players' level-k predicted payoffs have two possible values in treatment 3, due to different tie-breaking methods for the tie of a receiver classified as pooling or level-0. See footnote 5 for details.

come from level-1 senders with level-2 receivers, resulting in their payoffs to be $(-\frac{22}{3}, -\frac{10}{3})$, which are both worse than their nondisclosure counterparts. The *within-treatment* comparison is similar. In the nondisclosure subgame of treatment 3, the majority of interactions come from level-1 senders with pooling receivers, resulting in their payoffs to be $(-\frac{20}{3}, -\frac{8}{3})$. Meanwhile, in the disclosure subgame of treatment 3, the majority of interactions come from level-1 senders with level-2 receivers, resulting in their payoffs to be $(-\frac{22}{3}, -\frac{10}{3})$, which are both worse than their nondisclosure counterparts.

Thus, our level-k analysis suggests that some receivers tend to downgrade senders' exaggerated information more than necessary when the bias is disclosed, which results in a welfare loss to both parties and leads to lower average payoffs of both players compared with their nondisclosure counterparts.

7 Conclusion

We experimentally investigate the effect of disclosing private interest in the setting of strategic information transmission with unknown motives. A sender's interests are partially aligned with a receiver's and the sender has private information about his bias and the state. Our experiment provides support for the theory that mandatory disclosure of the private interest can be harmful to both senders and receivers. Moreover, the benefit of nondisclosure can only be realized when the private interest is automatically hidden. Meanwhile, our experimental data are inconsistent with the phenomena of *moral licensing* and *insinuation anxiety*, the psychological effects identified in the previous psychological studies of information disclosure with a direct conflict of interest, suggesting that these effects do not persist when the direct conflict of interest becomes partially aligned interests. We use a level-k model as an attempt to explain the payoff differences across different treatments and subgames and find that, according to our level-k model, the mismatch between senders and receivers with different levels of sophistication constitutes a source of welfare loss under bias disclosure. In particular, some receivers tend to downgrade senders' exaggerated information more than necessary when the bias is disclosed, which leads to a welfare loss to both parties and lower payoffs to both players compared with their nondisclosure counterpart.

Our experimental design could be extended in various directions. One possible way is to consider the setting in which receivers are ambiguous about senders' motives, by relaxing the assumption that receivers have perfect information about the distribution of senders' bias. Another extension could be the examination of private information disclosure in a repeated setting. This can be done, for example,

by fixing the groups of senders and receivers throughout the experiment. Our experiment could also be extended to a general distribution of senders' bias that has a nonzero mean, to capture the scenarios in which senders' private interests vary in both direction and degree and are intrinsically inclined towards a certain direction.

References

[1] Battaglini, M. (2002). Multiple referrals and multidimensional cheap talk. Econometrica, 70(4), 1379-1401.

[2] Battaglini, M., & Makarov, U. (2014). Cheap talk with multiple audiences: An experimental analysis. Games and Economic Behavior, 83, 147-164.

[3] Blume, A., Board, O. J., & Kawamura, K. (2007). Noisy talk. Theoretical Economics, 2(4), 395-440.

[4] Blume, A., DeJong, D. V., Kim, Y. G., & Sprinkle, G. B. (1998). Experimental evidence on the evolution of meaning of messages in sender-receiver games. The American Economic Review, 88(5), 1323-1340.

[5] Blume, A., Lai, E. K., & Lim, W. (2017). Strategic information transmission: A survey of experiments and theoretical foundations. Report.[1457].

[6] Cai, H., & Wang, J. T. Y. (2006). Overcommunication in strategic information transmission games. Games and Economic Behavior, 56(1), 7-36.

[7] Cain, D. M., Loewenstein, G., & Moore, D. A. (2005). The dirt on coming clean: Perverse effects of disclosing conflicts of interest. The Journal of Legal Studies, 34(1), 1-25.

[8] Chakraborty, A., & Harbaugh, R. (2010). Persuasion by cheap talk. American Economic Review, 100(5), 2361-2382.

[9] Chang, H., Chen, J., Duh, R., & Ittner, C. D. (2013). Do mandatory non-audit fee disclosures improve audit quality? Evidence from differential disclosure requirements. Drexel University. Working Paper.

[10] Chen, D. L., Schonger, M., & Wickens, C. (2016). oTree—An open-source platform for laboratory, online, and field experiments. Journal of Behavioral and Experimental Finance, 9, 88-97.

[11] Chen, Y., Kartik, N., & Sobel, J. (2008). Selecting Cheap-Talk Equilibria. Econometrica, 76(1), 117-136.

[12] Crawford, V. P., & Iriberri, N. (2007). Level-k auctions: Can a nonequilibrium model of strategic thinking explain the winner's curse and overbidding in private-value auctions?. Econometrica, 75(6), 1721-1770.

[13] Crawford, V. P., Costa-Gomes, M. A., & Iriberri, N. (2013). Structural models of nonequilibrium strategic thinking: Theory, evidence, and applications. Journal of Economic Literature, 51(1), 5-62.

[14] Crawford, V. P., & Sobel, J. (1982). Strategic information transmission. Econometrica: Journal of the Econometric Society, 1431-1451.

[15] Dimitrakas, V., & Sarafidis, Y. (2005). Advice from an expert with unknown motives.

[16] Farrell, J. (1993). Meaning and credibility in cheap-talk games. Games and Economic Behavior, 5(4), 514-531.

[17] Goltsman, M., Hörner, J., Pavlov, G., & Squintani, F. (2009). Mediation, arbitration and negotiation. Journal of Economic Theory, 144(4), 1397-1420.

[18] Goltsman, M., & Pavlov, G. (2011). How to talk to multiple audiences. Games and Economic Behavior, 72(1), 100-122.

[19] Gordon, S., Kartik, N., Lo, M. P. Y., Olszewski, W., & Sobel, J. (2023). Effective communication in cheap talk games.

[20] Hagenbach, J., & Perez-Richet, E. (2018). Communication with Evidence in the Lab. Games and Economic Behavior, 112, 139-165.

[21] Jin, G. Z., Luca, M., & Martin, D. (2021). Is no news (perceived as) bad news? An experimental investigation of information disclosure. American Economic Journal: Microeconomics, 13(2), 141-173.

[22] Kawagoe, T., & Takizawa, H. (2009). Equilibrium refinement vs. level-k analysis: An experimental study of cheap-talk games with private information. Games and Economic Behavior, 66(1), 238-255.

[23] Koch, C., & Schmidt, C. (2010). Disclosing conflicts of interest–Do experience and reputation matter?. Accounting, Organizations and Society, 35(1), 95-107.

[24] Lafky, J., Lai, E. K., & Lim, W. (2022). Preferences vs. strategic thinking: An investigation of the causes of overcommunication. Games and Economic Behavior, 136, 92-116.

[25] Li, M., & Madarasz, K. (2008). When mandatory disclosure hurts: Expert advice and conflicting interests. Journal of Economic Theory, 139(1), 47-74.

[26] Minozzi, W., & Woon, J. (2016). Competition, preference uncertainty, and jamming: A strategic communication experiment. Games and Economic Behavior, 96, 97-114.

[27] Morgan, J., & Stocken, P. C. (2003). An analysis of stock recommendations. RAND Journal of economics, 183-203.

[28] Ottaviani, M. (2000). The economics of advice. University College London, mimeo.

[29] Pei, H. D. (2015). Communication with endogenous information acquisition. Journal of Economic Theory, 160, 132-149.

[30] Rush, A., Smirnov, V., & Wait, A. (2009). Communication Breakdown: Consultation or Delegation from an Expert with Uncertain Bias. The BE Journal of Theoretical Economics, 10(1).

[31] Sachdeva, S., Iliev, R., & Medin, D. L. (2009). Sinning saints and saintly sinners: The paradox of moral self-regulation. Psychological science, 20(4), 523-528.

[32] Sah, S., Loewenstein, G., & Cain, D. M. (2013). The burden of disclosure: increased compliance with distrusted advice. Journal of personality and social psychology, 104(2), 289.

[33] Sah, S., Loewenstein, G., & Cain, D. (2019). Insinuation anxiety: Concern that advice rejection will signal distrust after conflict of interest disclosures. Personality and Social Psychology Bulletin, 45(7), 1099-1112.

[34] Shapiro, D., Shi, X., & Zillante, A. (2014). Level-k reasoning in a generalized beauty contest. Games and Economic Behavior, 86, 308-329.

[35] Sobel, J. (2020). Lying and deception in games. Journal of Political Economy, 128(3), 907-947.

[36] Vespa, E., & Wilson, A. J. (2016). Communication with multiple senders: An experiment. Quantitative Economics, 7(1), 1-36.

[37] Wang, J. T. Y., Spezio, M., & Camerer, C. F. (2010). Pinocchio's pupil: using eyetracking and pupil dilation to understand truth telling and deception in sender-receiver games. American economic review, 100(3), 984-1007.

[38] Weber, R. A., & Camerer, C. F. (2003). Cultural conflict and merger failure: An experimental approach. Management science, 49(4), 400-415.

Appendix A: Experimental Instructions - Treatment 1

INSTRUCTIONS

Welcome to this experiment. This experiment studies the interaction of decisions made by multiple individuals. In the following two hours or less, you will participate in **1** practice and **20** official rounds of decision making. Please read the instructions below carefully; the payment you will receive from this experiment depends on how well you make your decisions according to these instructions.

Your Role and Decision Group

There are **16** participants in today's session. One half of the participants will be randomly assigned the role of **Member A**, and the remaining one half of the participants the role of **Member B**. Your role will remain fixed throughout the experiment. Each group consists of one Member A and one Member B. The two members in a group make decisions that will affect their rewards in the round. Participants will be randomly rematched after each round to form new groups.

Your Decision in Each Round

In each round, the computer will randomly select a number among **1**, **3** and **5**. Each possible number has equal chance to be selected. The selected number will be revealed to Member A. Member B, without seeing the number, will have to make a guess. In the rest of the experiment, we will call the randomly selected number X.

Moreover, in each round, the computer will randomly select Member A's type that is either **HIGH** or **LOW**. Each possible type has equal chance to be selected. The selected type will be revealed to both Member A and Member B.

Member A privately learns X and makes a report to member B. Member B then makes a guess about X. **HIGH** type Member A wants Member B to make a higher guess, while **LOW** type Member A wants Member B to make a lower guess. We will explain it more in the later parts of the instructions.

Member A's Decisions

You will be presented with your type (**HIGH** or **LOW**) and the random number X. With all this information on your screen, you will be asked to report to Member B what X is. You do so by choosing a number from the three number boxes that represent 1, 3 and 5, after which you click the next button. You are free to choose any number box for your report; it is not part of the instructions that you have to tell the truth.

Once you click the next button, your decision in the round is completed and your report will be transmitted to your paired Member B, who will then be asked to make a guess.

Member B's Decisions

You will be presented with Member A's type (**HIGH** or **LOW**) and Member A's report about X. With all this information on your screen, you will be asked to make a guess about X by choosing a number from the five number boxes that represent 1, 2, 3, 4 and 5, after which you click the next button. You are free to choose any number box for your guess; it is not part of the instructions that you have to agree with the report.

Once you click the next button, your decision in the round is completed.

Your Reward in Each Round

Your reward in the experiment will be expressed in terms of points. The following describes how your reward in each round is determined.

Member A's Reward

The amount of points you earn in a round depends on your type, the random number X and Member B's guess. In particular,

If your type is **HIGH**, your reward $= 130 - 8 * [(X + 2) -$Member B's Guess $)]^2$,

which means that you are going to get the highest payoff if Member B makes the guess equal to $X + 2$. In case this value is negative, you will get 0.

If your type is **LOW**, your reward $= 130 - 8 * [(X - 2) -$Member B's Guess $)]^2$,

which means that you are going to get the highest payoff if Member B makes the guess equal to $X - 2$. In case this value is negative, you will get 0.

Member B's Reward

The amount of points you earn in a round depends on the random number X and your guess. In particular,

Your reward $= 130 - 8 * (X -$Member B's Guess $)^2$,

which means that you are going to get the highest payoff if your guess is the same as X. In case this value is negative, you will get 0.

The following table illustrates the payoffs for each player in different scenarios.

Random Number X	Member B's Guess	Member A's Reward		Member B's Reward
		HIGH Type	LOW Type	
1	1	98	98	130
	2	122	58	122
	3	130	2	98
	4	122	0	58
	5	98	0	2
3	1	2	130	98
	2	58	122	122
	3	98	98	130
	4	122	58	122
	5	130	2	98
5	1	0	98	2
	2	0	122	58
	3	2	130	98
	4	58	122	122
	5	98	98	130

Information Feedback

At the end of each round, the computer will provide a summary for the round: which number was selected and revealed to Member A, Member A's type, Member A's report, Member B's guess and your earnings in points.

Your Cash Payment

The experimenter randomly selects 1 round out of 20 official rounds to calculate your cash payment. (So it is in your best interest to take each round seriously.) Your total cash payment at the end of the experiment will be the amount of points you earned in the selected round plus a HK$40 show-up fee.

Quiz and Practice

To ensure your understanding of the instructions, we will provide you with a quiz and a practice round. We will go through the quiz after you answer it on your own.

You will then participate in 1 practice round. The practice round is a part of the instructions which is not relevant to your cash payment; its objective is to get you familiar with the computer interface and the flow of the decisions in each round.

Administration

Your decisions as well as your monetary payment will be kept confidential. Remember that you have to make your decisions entirely on your own; please do not discuss your decisions with any other participants.

Upon finishing the experiment, you will receive your cash payment. You will be asked to sign your name to acknowledge your receipt of the payment. You are then free to leave.

If you have any question, please raise your hand now. We will answer your question individually. If there is no question, we will proceed to the quiz.

Appendix B: Tables

Table 1: Equilibrium Predictions Under Bias Nondisclosure

Equilibrium	The Partition of Sender Types	Induced Actions
1	$\{H_1, H_3, H_5, L_1, L_3, L_5\}$	$\{3\}$
2	$\{\{H_1, L_5\}, \{H_3, H_5\}, \{L_1, L_3\}\}$	$\{3, 4, 2\}$
3	$\{\{H_3, H_5\}, \{H_1, L_1, L_3, L_5\}\}$	$\{4, 2\}$
4	$\{\{H_3, H_5\}, \{H_1, L_1, L_3, L_5\}\}$	$\{4, 3\}$
5	$\{\{L_1, L_3\}, \{H_1, H_3, H_5, L_5\}\}$	$\{2, 3\}$
6	$\{\{L_1, L_3\}, \{H_1, H_3, H_5, L_5\}\}$	$\{2, 4\}$
7	$\{\{H_1, L_1, L_3\}, \{H_3, H_5, L_5\}\}$	$\{2, 4\}$

Table 2: Non-Neologism-Proof Equilibria

Equilibrium	Self-Signaling Subset	Induced Action
1	$\{H_3, H_5\}$	4
3	$\{H_1, L_5\}$	3
4	$\{L_1, L_3\}$	2
5	$\{H_3, H_5\}$	4
6	$\{H_1, L_5\}$	3
7	$\{H_1, L_5\}$	3

Table 3: Best-Response Dynamics Analysis under Bias Nondisclosure

Players' Types	Strategies
Level-0 Left Sender	$L_1 = 1,\ L_3 = 3,\ L_5 = 5$
Level-0 Right Sender	$H_1 = 1,\ H_3 = 3,\ H_5 = 5$
Level-0 Receiver	$A_1 = 1,\ A_3 = 3,\ A_5 = 5$
Level-1 and above Left Sender	$L_1 = 1,\ L_3 = 1,\ L_5 = 3$
Level-1 and above Right Sender	$H_1 = 3,\ H_3 = 5,\ H_5 = 5$
Level-1 and above Receiver	$A_1 = 2,\ A_3 = 3,\ A_5 = 4$

Table 4: Best-Response Dynamics Analysis under Bias Disclosure: Left Sender

Players' Types	Strategies
Level-0 Sender	$L_1 = 1,\ L_3 = 3,\ L_5 = 5$
Level-0 Receiver	$A_1 = 1,\ A_3 = 3,\ A_5 = 5$
Level-1 Sender	$L_1 = 1,\ L_3 = 1,\ L_5 = 3$
Level-1 Receiver	$A_1 = 2,\ A_3 = 5$
Level-2 Sender	$L_1 = 1,\ L_3 = 1,\ L_5 = 1$
Level-2 Receiver	$A_1 = 3$

Table 5: Best-Response Dynamics Analysis under Bias Disclosure: Right Sender

Players' Types	Strategies
Level-0 Sender	$H_1 = 1,\ H_3 = 3,\ H_5 = 5$
Level-0 Receiver	$A_1 = 1,\ A_3 = 3,\ A_5 = 5$
Level-1 Sender	$H_1 = 3,\ H_3 = 5,\ H_5 = 5$
Level-1 Receiver	$A_3 = 1,\ A_5 = 4$
Level-2 Sender	$H_1 = 5,\ H_3 = 5,\ H_5 = 5$
Level-2 Receiver	$A_5 = 3$

Table 6: Sender's Expected Payoffs

	Disclosure	Nondisclosure
Level-0	-4	-4
Level-1	$-\frac{4}{3}$	$-\frac{4}{3}$
Level-2	$-\frac{11}{3}$	$-\frac{10}{3}$
Level-3 and above	$-\frac{20}{3}$	$-\frac{10}{3}$
Equilibrium	$-\frac{20}{3}$	$-\frac{10}{3}$

Table 7: Receiver's Expected Payoffs

	Disclosure	Nondisclosure
Level-0	0	0
Level-1	$-\frac{2}{3}$	-2
Level-2 and above	$-\frac{8}{3}$	-2
Equilibrium	$-\frac{8}{3}$	-2

Table 8: Disclosure/Detection Decisions

	Disclosure/Detection	Nondisclosure/Nondetection
Treatment 3	155(55.4%)	125(44.6%)
Treatment 4	277(92.3%)	23(7.7%)

Table 9: Subjects' Average Payoffs

	Sender	Receiver
Treatment 1	86.1	103.2
Treatment 2	88.9	107.7
Treatment 3, Aggregate	84.7	106.9
Treatment 3, Disclosure	86.4	107.2
Treatment 3, Nondisclosure	82.7	106.6
Treatment 4, Aggregate	83.1	106.3
Treatment 4, Detection	82.9	107.5
Treatment 4, Nondetection	86.1	91.7

Table 10: Subjects' Average Payoffs in the Last 10 Rounds

	Sender	Receiver
Treatment 1	78.7	99.8
Treatment 2	87.3	109.5
Treatment 3, Aggregate	84.3	104.1
Treatment 3, Disclosure	78.8	101.3
Treatment 3, Nondisclosure	89.4	106.8
Treatment 4, Aggregate	80.2	103.2

Table 11: Message Frequencies in Treatment 1

	Message=1	Message=3	Message=5
Negative Bias, State=1	79.59%	14.29%	6.12%
Negative Bias, State=3	80.95%	16.67%	2.38%
Negative Bias, State=5	29.79%	57.45%	12.77%
Positive Bias, State=1	12.96%	48.15%	38.89%
Positive Bias, State=3	6.38%	19.15%	74.47%
Positive Bias, State=5	2.44%	14.63%	82.93%

Table 12: Message Frequencies in Treatment 2

	Message=1	Message=3	Message=5
Negative Bias, State=1	75.93%	16.67%	7.41%
Negative Bias, State=3	82.98%	17.02%	0.00%
Negative Bias, State=5	0.00%	82.00%	18.00%
Positive Bias, State=1	19.67%	77.05%	3.28%
Positive Bias, State=3	0.00%	14.89%	85.11%
Positive Bias, State=5	4.92%	6.56%	88.52%

Table 13: Message Frequencies in the Disclosure Subgame of Treatment 3

	Message=1	Message=3	Message=5
Negative Bias, State=1	80.00%	20.00%	0.00%
Negative Bias, State=3	56.25%	40.63%	3.13%
Negative Bias, State=5	22.73%	50.00%	27.27%
Positive Bias, State=1	22.58%	58.06%	19.35%
Positive Bias, State=3	0.00%	40.74%	59.26%
Positive Bias, State=5	4.35%	0.00%	95.65%

Table 14: Message Frequencies in the Nondisclosure Subgame of Treatment 3

	Message=1	Message=3	Message=5
Negative Bias, State=1	77.27%	18.18%	4.55%
Negative Bias, State=3	82.35%	17.65%	0.00%
Negative Bias, State=5	0.00%	85.19%	14.81%
Positive Bias, State=1	10.53%	89.47%	0.00%
Positive Bias, State=3	0.00%	43.75%	56.25%
Positive Bias, State=5	0.00%	8.33%	91.67%

Table 15: Message Frequencies in the Detection Subgame of Treatment 4

	Message=1	Message=3	Message=5
Negative Bias, State=1	97.83%	2.17%	0.00%
Negative Bias, State=3	80.43%	19.57%	0.00%
Negative Bias, State=5	19.05%	61.90%	19.05%
Positive Bias, State=1	16.67%	66.67%	16.67%
Positive Bias, State=3	4.26%	25.53%	70.21%
Positive Bias, State=5	0.00%	9.52%	90.48%

Table 16: Action Frequencies in Treatment 1

	Action=1	Action=2	Action=3	Action=4	Action=5
Negative Bias, Message=1	33.33%	27.59%	35.63%	0.00%	3.45%
Negative Bias, Message=3	12.20%	0.00%	43.90%	12.20%	31.71%
Negative Bias, Message=5	0.00%	0.00%	20.00%	20.00%	60.00%
Positive Bias, Message=1	63.64%	0.00%	36.36%	0.00%	0.00%
Positive Bias, Message=3	24.39%	21.95%	43.90%	2.44%	7.32%
Positive Bias, Message=5	2.22%	0.00%	48.89%	22.22%	26.67%

Table 17: Action Frequencies in Treatment 2

	Action=1	Action=2	Action=3	Action=4	Action=5
Message=1	21.05%	33.68%	43.16%	2.11%	0.00%
Message=3	12.07%	6.90%	68.10%	9.48%	3.45%
Message=5	2.75%	0.00%	38.53%	36.70%	22.02%

Table 18: Action Frequencies in the Disclosure Subgame of Treatment 3

	Action=1	Action=2	Action=3	Action=4	Action=5
Negative Bias, Message=1	12.82%	25.64%	48.72%	10.26%	2.56%
Negative Bias, Message=3	10.71%	7.14%	57.14%	14.29%	10.71%
Negative Bias, Message=5	0.00%	0.00%	85.71%	0.00%	14.29%
Positive Bias, Message=1	37.50%	12.50%	37.50%	0.00%	12.50%
Positive Bias, Message=3	31.03%	10.34%	31.03%	27.59%	0.00%
Positive Bias, Message=5	2.27%	0.00%	52.27%	22.73%	22.73%

Table 19: Action Frequencies in the Nondisclosure Subgame of Treatment 3

	Action=1	Action=2	Action=3	Action=4	Action=5
Message=1	15.16%	27.27%	42.42%	12.12%	3.03%
Message=3	0.00%	10.71%	73.21%	14.29%	1.79%
Message=5	2.78%	0.00%	38.89%	41.67%	16.67%

Table 20: Action Frequencies in the Detection Subgame of Treatment 4

	Action=1	Action=2	Action=3	Action=4	Action=5
Negative Bias, Message=1	31.11%	18.89%	46.67%	2.22%	1.11%
Negative Bias, Message=3	5.56%	2.78%	25.00%	16.67%	50.00%
Negative Bias, Message=5	0.00%	0.00%	0.00%	25.00%	75.00%
Positive Bias, Message=1	90.91%	0.00%	9.09%	0.00%	0.00%
Positive Bias, Message=3	36.54%	13.46%	32.69%	11.54%	5.77%
Positive Bias, Message=5	2.50%	0.00%	48.75%	18.75%	30.00%

Table 21: Moral Licensing, Between-Treatment Comparison

Bias	State	Mean Difference_Disclosure	Mean Difference_Nondisclosure	Significance Level
Negative	1	2.53	2.63	None
Negative	3	0.43	0.34	None
Negative	5	0.85	0.36	0.005
Positive	1	1.04	0.46	0.005
Positive	3	0.64	0.30	0.1
Positive	5	2.39	2.33	None

Table 22: Moral Licensing, Within-Treatment Comparison

Bias	State	Mean Difference_Disclosure	Mean Difference_Nondisclosure	Significance Level
Negative	1	2.40	2.55	None
Negative	3	0.94	0.35	0.05
Negative	5	1.00	0.30	0.01
Positive	1	0.84	0.21	0.01
Positive	3	0.81	0.88	None
Positive	5	2.17	2.17	None

Table 23: Insinuation Anxiety, Between-Treatment Comparison

Message	Mean Difference_Disclosure	Mean Difference_Nondisclosure	Significance Level
1	1.08	1.26	0.05
3	0.94	0.47	0.005
5	1.22	1.25	None

Table 24: Insinuation Anxiety, Within-Treatment Comparison

Message	Mean Difference_Disclosure	Mean Difference_Nondisclosure	Significance Level
1	1.60	1.61	None
3	0.82	0.29	0.005
5	1.41	1.31	None

Table 25: Source of Disclosure, Treatment 1 v.s. The Disclosure Subgame of Treatment 3, Sender

Bias	State	Mean Difference_T1	Mean Difference_T3	Significance Level
Negative	1	2.53	2.40	None
Negative	3	0.43	0.94	0.05
Negative	5	0.85	1.00	None
Positive	1	1.04	0.84	None
Positive	3	0.64	0.81	None
Positive	5	2.39	2.17	0.1

Table 26: Source of Disclosure, Treatment 1 v.s. The Detection Subgame of Treatment 4, Sender

Bias	State	Mean Difference_T1	Mean Difference_T4	Significance Level
Negative	1	2.53	2.04	0.005
Negative	3	0.43	0.39	None
Negative	5	0.85	0.76	None
Positive	1	1.04	0.67	0.05
Positive	3	0.64	0.68	None
Positive	5	2.39	2.19	None

Table 27: Source of Disclosure, The Disclosure Subgame of Treatment 3 v.s. The Detection Subgame of Treatment 4, Sender

Bias	State	Mean Difference_T3	Mean Difference_T4	Significance Level
Negative	1	2.40	2.04	0.01
Negative	3	0.94	0.39	0.01
Negative	5	1.00	0.76	None
Positive	1	0.84	0.67	None
Positive	3	0.81	0.68	None
Positive	5	2.17	2.19	None

Table 28: Source of Disclosure, Treatment 1 v.s. The Disclosure Subgame of Treatment 3, Receiver

Bias	Message	Mean Difference_T1	Mean Difference_T3	Significance Level
Negative	1	1.13	1.64	0.005
Negative	3	1.00	0.64	0.1
Negative	5	0.60	1.71	0.01
Positive	1	0.73	1.38	None
Positive	3	0.88	1.00	None
Positive	5	1.29	1.36	None

Table 29: Source of Disclosure, Treatment 1 v.s. The Detection Subgame of Treatment 4, Receiver

Bias	Message	Mean Difference_T1	Mean Difference_T4	Significance Level
Negative	1	1.13	1.23	None
Negative	3	1.00	1.31	0.1
Negative	5	0.60	0.25	None
Positive	1	0.73	0.18	0.1
Positive	3	0.88	1.10	None
Positive	5	1.29	1.26	None

Table 30: Source of Disclosure, The Disclosure Subgame of Treatment 3 v.s. The Detection Subgame of Treatment 4, Receiver

Bias	Message	Mean Difference_T3	Mean Difference_T4	Significance Level
Negative	1	1.64	1.23	0.05
Negative	3	0.64	1.31	0.005
Negative	5	1.71	0.25	0.005
Positive	1	1.38	0.18	0.01
Positive	3	1.00	1.10	None
Positive	5	1.36	1.26	None

Table 31: Level-k Classification

Sender, Disclosure	0, 1, 2
Sender, Nondisclosure	0, 1 (Equilibrium)
Receiver, Disclosure	0, 1, 2, Pooling
Receiver, Nondisclosure	0, 1 (Equilibrium), Pooling

Table 32: Level-k Classification in Treatment 1

	Sender	Receiver
Level-0	1(7.14%)	2(14.29%)
Level-1	8(57.14%)	1(7.14%)
Level-2	4(28.57%)	3(21.43%)
Pooling	–	2(14.29%)
Unclassified	1(7.14%)	6(42.86%)
Total	14	14

Table 33: Level-k Classification in Treatment 2

	Sender	Receiver
Level-0	1(6.25%)	1(6.25%)
Level-1	14(87.50%)	5(31.25%)
Pooling	–	5(31.25%)
Unclassified	1(6.25%)	5(31.25%)
Total	16	16

Table 34: Level-k Classification in Treatment 3, Disclosure

	Sender	Receiver
Level-0	4(28.57%)	0(0.00%)
Level-1	7(50.00%)	1(7.14%)
Level-2	1(7.14%)	5(35.71%)
Pooling	–	1(7.14%)
Unclassified	2(14.29%)	7(50.00%)
Total	14	14

Table 35: Level-k Classification in Treatment 3, Nondisclosure

	Sender	Receiver
Level-0	1(11.11%)	1/2(7.14%/14.29%)
Level-1	8(88.89%)	3(21.43%)
Pooling	–	6/5(42.86%/35.71%)
Unclassified	0(0.00%)	4(28.57%)
Total	9	14

Table 36: Level-k Classification in Treatment 4

	Sender	Receiver
Level-0	2(13.33%)	2(13.33%)
Level-1	11(73.33%)	2(13.33%)
Level-2	2(13.33%)	6(40.00%)
Pooling	–	0(0.00%)
Unclassified	0(0.00%)	5(33.33%)
Total	15	15

Table 37: Payoff Matrix under Bias Disclosure

	Level-0 Receiver	Level-1 Receiver	Level-2 Receiver	Pooling Receiver
Level-0 Sender	$(-4, 0)$	$(-9, -\frac{11}{3})$	$(-8, -4)$	$(-\frac{20}{3}, -\frac{8}{3})$
Level-1 Sender	$(-\frac{4}{3}, -\frac{8}{3})$	$(-\frac{14}{3}, -\frac{2}{3})$	$(-\frac{22}{3}, -\frac{10}{3})$	$(-\frac{20}{3}, -\frac{8}{3})$
Level-2 Sender	$(-\frac{8}{3}, -\frac{20}{3})$	$(-\frac{11}{3}, -\frac{11}{3})$	$(-\frac{20}{3}, -\frac{8}{3})$	$(-\frac{20}{3}, -\frac{8}{3})$

Table 38: Payoff Matrix under Bias Nondisclosure

	Level-0 Receiver	Level-1 Receiver	Pooling Receiver
Level-0 Sender	$(-4, 0)$	$(-\frac{14}{3}, -\frac{2}{3})$	$(-\frac{20}{3}, -\frac{8}{3})$
Level-1 Sender	$(-\frac{4}{3}, -\frac{8}{3})$	$(-\frac{10}{3}, -2)$	$(-\frac{20}{3}, -\frac{8}{3})$

Table 39: Subjects' Level-k Predicted Payoffs

	Sender	Receiver
Treatment 1	86.54	105.92
Treatment 2	92.21	111.54
Treatment 3	81.45/78.38	107.59/107.48
Treatment 4	83.64	107.10