Knowing More Can Hurt: An Experiment*

Weixuan Zhou†

November 10, 2023

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 and insinuation anxiety. 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 a unique babbling equilibrium, whereas informative equilibria exist when the private interest is hidden. We then use neologism-proofness 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. We perform a level-k analysis to explain this phenomenon and find that a source of the welfare loss when private interest is disclosed is the mismatch between senders and receivers with different levels of sophistication. Meanwhile, our experimental data are inconsistent with the phenomena of moral licensing and insinuation anxiety, suggesting that these psychological effects do not persist when direct conflict of interest becomes partially aligned interests.

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

JEL Codes: C91, D82, D83

*The author is deeply indebted to Wooyoung Lim for his detailed comments, guidance and help. The author thanks Pak Hung Au, Rui Tang, Qinggong Wu, Xu Zhang and participants of the 2023 Stony Brook International Conference in Game Theory 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

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.

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. Through a direct channel, it reduces the information gap between the informed and the uninformed (Healy and Palepu, 2001, p.412). There is also a more subtle channel through which disclosure of private interest can at least benefit the expert. 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).

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. Denote the sender as high type if his bias is positive and low type 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 our first treatment, the private interest is always disclosed. In our second treatment, the private interest is automatically hidden. In our third treatment, the sender decides whether to disclose his bias to the receiver before observing the state and his bias. In our 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 effect across different sources of information disclosure. Each treatment consists of two identical sessions.

Our theoretical prediction shows that nondisclosure of private interest can facilitate information transmission and benefit both parties. Intuitively, it creates a possibility for a high type sender who observes a low state to pool with a low type sender who observes a high state. In this case, an informative message is transmitted and the receiver perfectly identifies the true state. 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, only the babbling equilibrium (or those essentially equivalent ones) can be realized.

Since multiple equilibria arise when the private interest is hidden, 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 that is realized when the private interest is disclosed. As for the best-response dynamics approach, we adopt a level-k framework that following Crawford and Iriberri (2007). We first specify a level-0 sender to be truthtelling. Then, a level-k receiver best responds to a level-k sender and a level-k sender best responds to a level-(k-1) receiver. Our level-k model converges to the same sender-optimal equilibrium 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* comparison, and the payoff differences are particularly stark 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 difference in average payoffs across treatments is even larger in the last 10 rounds, when players stabilize their strategies. 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. Moreover, between the two sessions of treatment 3, both senders and receivers achieve a higher payoff in the session where bias disclosure is less frequent.

To provide a systematic explanation of the observed payoff differences, we perform a level-k analysis and characterize players into different levels of sophistication according to their strategies. We find that a source of welfare loss when private interest is disclosed is the mismatch of senders and receivers with different levels of sophistication. Meanwhile, our experimental data are inconsistent with the phenomena of moral licensing and insinuation anxiety, suggesting that these psychological effects do not persist when direct conflict of interest becomes partially aligned interests.

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. Section 7 concludes. Experimental instructions are rendered in the appendix.

2 Literature Review

Our experiment follows closely from the theoretical work of Li and Madarasz (2008) on strategic information transmission with unknown motives. There are a few theoretical studies in this direction. For example, 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.

A few experimental and empirical studies on strategic information transmission with unknown motives have also been performed. Cain et. al. (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 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 is substantial difference. 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 paper is also related to applications of the best-response dynamics approach. To name a few, Cai and Wang (2005) study the phenomenon of overcommunication in a cheap-talk game, Crawford and Iriberri (2007) study the phenomenon of overbidding in an auction model, and Shi and Zillante (2014) study a class of generalized beauty contests.

3 Theoretical Background

Our experimental design is motivated by the leading example of Li and Madarasz (2008), originating from Crawford and Sobel (1982), the seminal work in strategic information transmission.

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. 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 ideal actions of the players, 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. All players 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 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 treatment, the receiver knows only the distribution of the sender's bias. 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 treatment, the sender can choose whether to reveal his bias to the receiver before knowing his bias and the state. In the last treatment, the receiver can choose whether to detect the sender's bias before the sender knows his bias and the state.

In our experiment, treatment 1 and treatment 2 serve as the benchmark cases. Treatment 3 consists of two subgames: 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: if the receiver chooses to detect the sender's bias, players reach the detection subgame; otherwise, players reach the nondetection subgame. Figure 1 and 2 show the game structures of treatment 3 and 4, respectively.

Figure 1: The Game Structure of Treatment 3

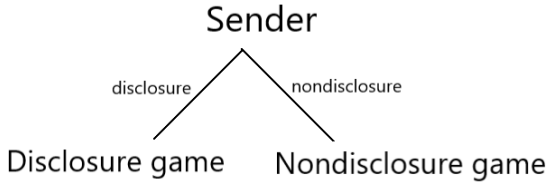
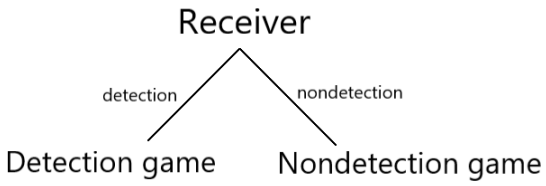


Figure 2: The Game Structure of Treatment 4



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 under both bias disclosure and bias nondisclosure. Treatment 1, the disclosure subgame of treatment 3 and the detection subgame of treatment 4 correspond to bias disclosure, whereas treatment 2, the nondisclosure subgame of treatment 3 and the nondetection subgame of treatment 4 correspond to bias nondisclosure.

4.1 Equilibrium Predictions 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 $\{\{1,3,5\}\}$. Among them, only $\{\{1,5\},\{3\}\}$ and $\{\{1,3,5\}\}$ constitute an equilibrium. Note that the receiver will ignore the message and choose $y = 3$ in both partitions, which means that babbling equilibrium is the essentially unique equilibrium outcome under bias disclosure. Each sender's expected payoff is $-\frac{20}{3}$, and the receiver's expected payoff is $-\frac{8}{3}$. Proposition 1 summarizes our finding under bias disclosure.

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

4.2 Equilibrium Predictions under Bias Nondisclosure

An equilibrium is characterized by a partition of the product space of the state space and the distribution of the sender's bias, which can be denoted as $T = \{L_1, L_3, L_5, R_1, R_3, R_5\}$. For example, the partition $\{\{L_1\},\{L_3, L_5, R_1\},\{R_3, R_5\}\}$ corresponds to an equilibrium in which the left sender sends m_1 when the state is 1 and m_2 when the state is 3 and 5, the right sender sends m_2 when the state is 1 and m_3 when the state is 3 and 5, and the receiver optimally responds by choosing 1, 3 and 4 upon receiving m_1, m_2 and m_3, respectively. In any equilibrium, different types of the sender who induce the same action for the receiver are pooled together. It turns out that the game has seven (essentially distinct) equilibria, which are summarized in Table 1.

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	$\{\{H_1, L_1, L_3\}, \{H_3, H_5, L_5\}\}$	$\{2, 4\}$

Among the equilibria, equilibrium 2 and 7 are Pareto optimal in the sense that any other equilibrium is Pareto inferior to them. More specifically, equilibrium 2 is sender optimal and equilibrium 7 is receiver optimal. This result is slightly different from the one in Li and Madarasz (2008), in which all equilibria are Pareto ranked. The difference arises from the structure of the state space, which is finite and discrete in our setting but is continuous in Li and Madarasz (2008). Proposition 2 summarizes our equilibrium predictions under bias nondisclosure.

Proposition 2 *Under bias nondisclosure, multiple equilibria exist. Among them, there exists a sender-optimal equilibrium and a receiver-optimal equilibrium.*

4.3 Equilibrium Selection under Bias Nondisclosure

To sharpen our theoretical prediction and address the issue of multiple equilibria, we perform equilibrium selection using both neologism-proofness (Farrell, 1993) and best-response dynamics analysis (Crawford and Iriberri (2007)). Both approaches select the sender-optimal equilibrium as the unique outcome. [2]

4.3.1 Neologism-Proofness

In this part, we use the concept of neologism-proofness according to Farrell (1993) to select an equilibrium. For any equilibrium, define $T_S \subset T$ as a self-signaling subset if any sender of type $t \in T_S$

[2]Qualitatively, our theoretical results hold when $\frac{3}{2} \leq b \leq \frac{5}{2}$ in the sense that the sender optimal equilibrium continues to exist, survives intuitive criterion and continues to be the only converging outcome of our best response dynamics approach.

is strictly better off when the receiver acts optimally according to T_S than according to the equilibrium and any other type that does not belong to T_S does not want to induce that action instead of 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.

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

Proposition 3 *Under bias nondisclosure, only the sender-optimal equilibrium is neologism-proof.*

4.3.2 Best-Response Dynamics Analysis

In this part, we perform best-response dynamics analysis using a level-k model. We assume that each player can be classified according to her level of sophistication, denoted as a level-k sender or a level-k receiver. Players' strategies can be iteratively determined once the strategies of level-0 senders are specified. In particular, 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 $j \in \{1, 3, 5\}, L_j \subseteq \{1, 3, 5\}, H_j \subseteq$

$\{1,3,5\}, A_j \subseteq \{1,2,3,4,5\}$. Table 3 summarizes the best-response dynamics analysis. Proposition 4 summarizes our finding.

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	$A_1 = 2,\ A_3 = 3,\ A_5 = 4$

Proposition 4 *Under bias nondisclosure, according to 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.

Theorem 1 *Under bias disclosure, 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 convergence outcome of the best-response dynamics approach.*

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; no actions were taken in treatment 1 or 2. Then, in all treatments, the sender privately observed the state θ and his type of bias (i.e., whether he was a left sender or a right sender). After that, in treatment 1, the sender was obliged to disclose his bias to the receiver, whereas in treatment 2 the bias was automatically hidden. 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 message, the receiver's action and the subject's own payoff.

5.2 Hypotheses

To postulate whether players would choose to disclose/detect the bias or not and to compare their expected payoffs across treatments, we calculate players' expected payoffs given their level of sophistication under bias disclosure and nondisclosure. More specifically, we assume for each $k \geq 1$, a level-k sender's expected payoff is calculated according to her optimal strategies when she interacts with a level-(k-1) receiver and a level-k receiver's expected payoff is calculated according to his strategies when he interacts with a level-k sender. Finally, we assume a level-0 sender's expected payoff is 0 by assuming that she interacts with a credulous sender who takes an action that is always 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 players' expected payoffs in different circumstances.

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

Our first hypothesis concerns the equilibrium predictions in treatment 1 and 2. We formulate this hypothesis based on our results 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 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. This is because both senders and receivers are better off in the sender-optimal equilibrium under bias nondisclosure than in the babbling equilibrium under bias disclosure. Our alternative hypothesis is the converse, which may arise from psychological effects such as moral licensing and insinuation anxiety.

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' decisions on bias disclosure/ detection and their average payoffs within and across treatments. In section 6.2, we perform a level-k analysis to explain the variation in subjects' earnings. We characterize subjects into different levels of sophistication according to their behavior and calculate their predicted payoffs based on the empirical level-k distribution. We find that a source of welfare loss when private

interest is disclosed is the mismatch of senders and receivers with different levels of sophistication. In section 6.3, we compare our experimental data with the predictions based on the psychological effects. We find that our experimental data are inconsistent with the phenomena of moral licensing and insinuation anxiety, suggesting that these psychological effects do not persist when direct conflict of interest becomes partially aligned interests.

6.1 Disclosure/Detection Decisions and Payoffs

Table 8 summarizes senders' decisions on bias disclosure in treatment 3 and receivers' decisions on bias detection in treatment 4. 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 8: Disclosure/Detection Decision

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

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, but this finding is subject to small sample bias. Table 9 shows that more than 90% of the observations in treatment 4 fall into bias detection and only less than 10% of the observations in treatment 4 fall into nondetection. Moreover, in one session of treatment 4, all receivers choose to detect the sender's bias in the last 10 rounds.

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

The payoff differences are more quantitatively and statistically significant in the last 10 rounds. Table 10 summarizes subjects' average payoffs across treatments in the last 10 rounds. Players' average payoffs across the four treatments exhibit a clear pattern, 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 is statistically significant at the 0.05 level between treatment 1 and treatment 2 (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).

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

Player-level data also provide evidence for the finding. In treatment 3 in the last 10 rounds, two

out of seven senders always disclose and two out of seven senders never disclose. The average payoffs of nondisclosing senders are 87.6 and 79.6, whereas the average payoffs of disclosing senders are 58.0 and 53.2, and the difference is statistically significant at the 0.05 significance level (p=0.04, Wilcoxon signed-rank test).

6.2 Level-k Analysis

Previous experiments in strategic information transmission have shown that subjects do not necessarily conform to equilibrium strategies, but instead systematically depart from them. 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 babbling under bias disclosure and as level-0, level-1/ equilibrium or babbling under bias nondisclosure. [3] 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 fall into this group. Table 11 summarizes our classification method. Based on our 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. [4]

Table 11: Level-k Classification

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

Tables 12-16 summarize our level-k classification. Overall, more than 90% of the senders can be

[3] 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.

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

classified into at least one category (94.12%), while the fraction of receivers is 63.01%. Intuitively, a sender have all the relevant information of the state, his bias and his ideal action, and thus is easier to determine his communication strategies, whereas a receiver is faced with multidimensional uncertainty and multiple decision makings (in treatment 4). Across all the treatments, most senders are classified as level-1 (70.59%), suggesting that senders typically choose a message closest to their ideal actions. The classification patterns of receivers, meanwhile, have more variations. Under bias disclosure, most receivers are classified as level-2 (56.00%), while under bias nondisclosure, most are classified as either babbling (52.38%/47.62%) or level-1 (38.10%).

Table 12: 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%)
Babbling	NA	2(14.29%)
Unclassified	1(7.14%)	6(42.86%)
Total	14	14

Table 13: Level-k Classification in Treatment 2

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

Table 14: 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%)
Babble	NA	1(7.14%)
Unclassified	2(14.29%)	7(50.00%)
Total	14	14

Table 15: 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%)
Babble	NA	6/5(42.86%/35.71%)
Unclassified	0(0.00%)	4(28.57%)
Total	9	14

Table 16: 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%)
Babble	NA	0(0.00%)
Unclassified	0(0.00%)	5(33.33%)
Total	15	15

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 occur 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 1/5 from a right/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 assumptions on off-equilibrium strategies when classifying subjects. Table 17 and 18 summarize our results. In each vector, the first entry indicates the sender's payoff and the second entry indicates the receiver's payoff.

Table 17: Payoff Matrix under Bias Disclosure

	Level-0 Receiver	Level-1 Receiver	Level-2 Receiver	Babble 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 18: Payoff Matrix under Bias Nondisclosure

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

Next, we calculate subjects' level-k predicted payoffs according to the empirical distribution of subjects in each treatment and scale up the payoffs to match the payoff functions in the actual experiment. Table 19 summarizes our results.

Table 19: 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

The level-k predicted payoffs in Table 19 closely match the actual payoffs in Table 10, suggesting that our level-k classification works reasonably well in explaining players' observed behaviors. Tables 12-18, together, provide 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 babbling. 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 babbling 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 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 babble receivers in treatment 2, resulting their payoffs to be $(-\frac{10}{3}, -2)$ or $(-\frac{20}{3}, -\frac{8}{3})$. Meanwhile, under bias disclosure, the majority of interactions come from level-1 senders with level-2 receivers, resulting their payoffs to be $(-\frac{22}{3}, -\frac{10}{3})$, which are both worse than their nondisclosure counterparts as level-2 receivers downgrade the level-1 sender's information more than necessary. The *within-treatment* comparison is similar. In the nondisclosure subgame of treatment 3, the majority of interactions come from level-1 senders with babble receivers, resulting 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 their payoffs to be $(-\frac{22}{3}, -\frac{10}{3})$, which are both worse than their nondisclosure counterparts.

6.3 Moral Licensing, Insinuation Anxiety and Source of Disclosure

Studies in psychology have shown that disclosing private interest can result in the phenomena of *moral licensing* and *insinuation anxiety* (Cain, Loewenstein and Moore, 2005 and 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. It is important to note that these studies are under the setting that senders have a direct conflict of interest with receivers in the form of a zero-sum game. We would like to know whether similar effects persist in our setting, where senders' interests are partially aligned with receivers'.

Our analysis on *moral licensing* consists of both *between-treatment* and *within-treatment* comparison. 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 can choose whether to disclose their bias.

Following the spirit of Cain, Loewenstein and Moore (2005) and Sah, Loewenstein and Cain (2018), which show that the source of disclosure (e.g., external disclosure or disclosure by senders) also affects players' actions and payoffs, we also analyze the effect of source of disclosure in our setting. To do so, we compare senders' data in treatment 1, the disclosure subgame of treatment 3 and the detection subgame in treatment 4, and compare receivers' data in treatment 2 and the nondisclosure subgame of treatment 3. [5]

Tables 20-29 summarize the frequencies of senders' messages and receivers' actions across different treatments/subgames. Tables 20 and 21 demonstrate whether the *moral licensing* effect persists from the *between-treatment* comparison. The first two lines 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 when the state is 1 and choose message 5 more frequently when the state is 3 in the disclosure subgame of treatment 3. 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*.

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

Tables 22 and 23 demonstrate whether the *moral licensing* effect persists from the *within-treatment* comparison. With negative bias, senders choose message 1 more frequently when the state is 3 and choose message 3 more frequently when the state is 5 in the nondisclosure subgame of treatment 3. With positive bias, senders choose message 3 more frequently when the state is 1 in the nondisclosure subgame. Our *within-treatment* comparison also suggests that senders' messages are closer to their ideal action under bias nondisclosure.

Tables 20, 22 and 24 demonstrate whether the source of private information disclosure plays a role in senders' communication strategies. Comparing Table 20 with Table 22, 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 20 with Table 24, 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 22 with Table 24, 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 mandatory bias disclosure.

Tables 25, 27 and 29 demonstrate whether the source of private information disclosure plays a role in receivers' actions. Comparing Table 25 with Table 27, 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 25 and Table 29 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 27 and Table 29, we find that receivers choose action 1 more frequently when the bias is positive or negative and the message is 1, 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 that relates the source of disclosure to receivers' action profiles.

Table 20: 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 21: 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 22: 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 23: 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 24: 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 25: 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 26: 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 27: 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 28: 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 29: 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%

On top of the summary statistics, we also provide formal test statistics on these comparisons. Tables 30-39 summarize our test results. Table 30 and Table 31 present the Wilcoxon signed-rank test statistics of *between-treatment* and *within-treatment* comparisons of senders' communication strategies. We compute the differences between senders' messages and their ideal actions for each state, and compare the differences between disclosure and nondisclosure. Table 30 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 bias disclosure, and there are no statistically significant differences in other cases. Table 31 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 bias disclosure, and there are no statistically significant differences in other cases. Therefore, tables 30-31, together, suggest that the senders only exhibit the psychological effect of *moral licensing*, if at all, when their private interest is hidden.

Tables 32 and 33 present the Wilcoxon signed-rank test test statistics of *between-treatment* and *within-treatment* comparisons of receivers' communication strategies. 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 opposite 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, our test results are inconsistent with the psychological effect of *insinuation anxiety*.

Tables 34-36 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 34 shows that senders choose a message closer to their ideal actions when the bias is negative and the state is 3 in treatment 1, but the converse is true when the bias is positive and the state is 1. Table 35 and Table 36 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 37-39 present the Wilcoxon signed-rank test statistics of receivers' actions against different sources of disclosure. Table 37 shows that receivers take an action that is closer to senders' messages 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 38 shows that receivers take an action that is closer to senders' messages 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 39 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 of *moral licensing* and *insinuation anxiety*, suggesting that these psychological effects do not persist when the direct conflict of interest becomes partially aligned interests.

Table 30: 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 31: 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 32: 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 33: 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 34: 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 35: 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 36: 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 37: 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 38: 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 39: 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

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 have private information about his bias and the state. Our experiment provides support for the theory that mandatory disclosure of 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. Using a level-k model, we find that the mismatch between senders and receivers of different levels of sophistication constitutes a source of welfare loss under bias disclosure. Meanwhile, our experimental data are inconsistent with the phenomena of moral licensing and insinuation anxiety, the psychological effects identified in previous experimental studies of information disclosure with direct conflict of interests. Therefore, our experiment suggests that these psychological effects do not persist when the direct conflict of interest becomes partially aligned interests.

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

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

[3] 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.

[4] 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.

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

[6] 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.

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

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

[9] 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.

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

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

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

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

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

[15] 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).

[16] 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.

[17] 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.

[18] 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.

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

Appendix: 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.