Confusing Context with Character: Correspondence Bias in Economic Interactions

Yi Han, Yiming Liu, and George Loewenstein*

Abstract

When drawing inferences about a person's personal characteristics from their actions, "correspondence bias" is the tendency to overestimate the influence of those characteristics and underestimate the influence of situational factors, such as incentives the individual faces. We build a simple framework to formalize correspondence bias, and test its predictions in an online experiment. Consistent with correspondence bias, subjects are, on average, willing to pay to receive the dictator-game givings of an individual with whom they are randomly assigned to play a game that encourages cooperation rather than one with whom they play a game that encourages selfish behavior. We show, further, that experiencing both games oneself, as opposed to playing one and observing the other, reduces the bias, and receiving information about how each of the players behaved in both games, eliminates it.

Key Words: Belief Updating, Attribution Bias, Incentives, Experiments

JEL Classifications: C91, D90

---

*Han: School of Applied Economics, Renmin University of China, yihanecon@ruc.edu.cn; Liu: Humboldt University of Berlin, WZB Berlin Social Science Center, yiming.liu@wzb.eu; Loewenstein: Department of Social and Decision Sciences, Carnegie Mellon University, gl20@andrew.cmu.edu. We thank Kareem Haggag, David Huffman, Lise Vesterlund, Alistair Wilson for their helpful comments. We also thank the departmental editor, Yan Chen, and two anonymous referees for their helpful and constructive feedback.

1 Introduction

When drawing inferences about a person's enduring characteristics from their behaviors, the *correspondence bias* (Jones and Harris, 1967; Ross, 1977; Gilbert and Malone, 1995) is the tendency to overestimate the influence of the person's enduring characteristics on decisions they make, and to underestimate the impact of situational factors, such as social pressures.

Correspondence bias, which was previously overlooked by economists, has important implications for interpersonal interactions. When employers decide whom to hire, when college admission officers decide whom to admit, and when a board of directors decides how highly to remunerate a CEO, they usually cannot directly observe the qualities of the relevant entity, but need to make inferences about them from their observed behaviors and outcomes. Complicating this process is the fact that people's choices are also heavily influenced by situational factors such as, in the examples just provided, the challenges an employee faced in their last job, whether the high school student went to a school that grades leniently or strictly, and whether the CEO was hired just before the sector their company is in performed well or poorly. In settings such as these, a fully rational Bayesian decision-maker would be able to disentangle the influences of different incentives on behaviors and outcomes to back out an unbiased guess of an individual's underlying characteristics. A correspondence-biased decision maker would, in contrast, systematically underestimate the impact of situational factors that people face, and so over-attribute behaviors and outcomes to the decision makers' characteristics, such as work ethic and intelligence.

Among the situational factors that can influence an individual's behavior, *incentives* are of special interest to economists. Both when it comes to individual-level decision making and interpersonal strategic behavior, research documenting impacts of incentives on behavior is so extensive as to defy systematic review. However, as just one example of particular relevance to the research reported herein, in economic games for which defection is a dominant strategy, prior research has found that people cooperate more when the payoff from mutual cooperation is higher (Charness et al., 2016), when "punishment" from cooperating unilaterally is smaller, and when the payoff from defecting against a cooperator is lower (Mengel, 2018). In this setting, a Bayesian decision maker who wanted to explain an individual's behavior in a specific game would make unbiased judgments of the individual's characteristics, effectively controlling for the game they are observed playing; in contrast a correspondence-biased decision maker would over-attribute the player's action to their characteris-

tics and fail to adjust their judgments sufficiently for the impact of the game the observed agent was playing.

We build a simple framework to formalize the idea of correspondence bias. In our model, an individual chooses between two players after observing their actions, and the goal is to choose the one who is more likely to be the Good type. One player plays the *benign game* in which both the Good type and the Bad type choose to cooperate in equilibrium, while the other player plays the *malign game* in which the Good type cooperates and the Bad type defects in equilibrium. Borrowing the idea of *cursed equilibrium* (Eyster and Rabin, 2005), we model correspondence bias as the tendency to underestimate the correlation between actions and the game structure when interpreting information obtained about others' play. As cooperation in the malign game is a strong signal of the Good type and cooperation in the benign game is a weak signal of the Good type, confusing between the two games leads the biased individual to over-interpret cooperation in the benign game as a good signal, and under-interpret cooperation in the malign game as a good signal. Even when the game assignment is completely determined by chance, our model predicts that a correspondence-biased individual is willing to incur a cost to choose the benign-game player.

There are several challenges to empirically identifying correspondence bias. Imagine an experiment in which we randomly assign half of the subjects to play the benign game that incentivizes everyone to cooperate, and the other half to play the malign game that incentivizes some people to defect. We then let them choose between a benign-game player and a malign-game player to play a follow-up game together. Our model predicts that, in expectation, people are willing to pay more for the former. The first challenge is to distinguish between correspondence bias and Bayesian updating. Choosing the benign-game player can be consistent with Bayesian updating, as someone who chooses to cooperate in the benign game can rationally be expected to be more prosocial than someone who chooses to defect in the malign game. Second, reciprocity can also motivate choosing the benign game player; subjects may want to reciprocate the benign-game player's cooperative behavior in the follow-up game. Third, if one believes that the games have behavioral spillover effects on people's prosociality, and specifically that playing good (bad) games makes people more (less) prosocial, as shown in Bednar et al. (2012); Peysakhovich and Rand (2015) and Cason et al. (2019), then it makes sense to choose the individual who played the benign game.

We seek to rule out these three potential confounds using a three-stage experimental design. In

the first stage, all subjects make a decision as the dictator in the dictator game. In the second stage, they are randomly matched into groups of four to play the benign game and the malign game. Both games are 2x2 complete-information games with a strictly dominant strategy for both players. The malign game is the classic prisoner's dilemma game in which the dominant strategy is to defect, while the benign game is the Harmony Game (Dal Bó et al., 2018) in which the dominant strategy is to cooperate. At the end of the second stage, subjects are able to see the actions of one or more players, and to obtain information about the payoff structure of the games they played. Based on this information, in the third stage, they choose which of two players to receive the dictator givings from, and we use a multiple price list to elicit their willingness to pay (WTP) for their preferred player.

We address the Bayesian updating confound by randomly assigning players to the two games. This randomization ensures that the benign-game players and malign-game players are equally likely to be the Good type *ex ante*. The Martingale property of Bayesian beliefs then implies that the expected posterior beliefs are the same; a Bayesian model predicts that the individual will be in expectation indifferent between receiving the dictator offerings of the two players. However, our model predicts that a correspondence-biased individual will be (in expectation) willing to pay a positive amount to be matched with the benign-game player.

Our design avoids the possibility that reciprocity could drive the results by using a dictator game in which there are no actions that the receiver can take; thus, there is no way to reciprocate the benign-game player's cooperation in the follow-up game.[1]

Finally, we avoid the potential for positive behavioral spillover from participation in the benign game by sequencing the dictator decision so it occurs *before* Stage 2, the stage when subjects play the benign and the malign games. Even if individuals become more prosocial after playing the benign game, the dictator decision will have already been made in Stage 1, and cannot be altered by the game.

In the baseline treatment, Treatment 2, subjects only play one game, but those who played the benign (malign) game also learn about the action of a malign-game (benign-game) player at the end

[1] In addition, subjects cannot reciprocate the benign-game player through choosing them as the dictator either. Due to the matching protocol in Stage 3, the experiment is set up so that choosing a player as the dictator does not benefit him/her.

of Stage 2. They are informed about the payoffs of the game played by the other player, as well as the other player's action, but they do not experience the game themselves. In Stage 3 they choose whether to obtain the dictator-game givings of the benign-game player or the malign-game player.

Our results show, first, that correspondence bias exists and influences Stage 3 decisions. We measure the impact of correspondence bias through the *benign premium* – the extra amount a subject, in Stage 3, is willing to pay for the dictator game givings of a benign-game player compared to the dictator-game givings of a malign-game player, which the players had decided upon in Stage 1. While the rational Bayesian model predicts the benign premium to be 0, we find that the benign premium is 11.67 cents on average in Treatment 2, the baseline treatment, which is significantly different from 0 at the 1% level. To receive the dictator game givings of the player who is randomly assigned to the benign game, subjects are on average willing to give up 6% of the $2.00 divided by the dictator (which is the largest possible difference between the two potential dictators), or 12% of the $1.00 (half of the 'pie' is the typical modal amount given in the dictator game; only 11 out of 817 subjects, or 1 percent of subjects, in our experiment gave more than $1.00).

To understand the mechanism behind correspondence bias, and to explore potential methods to reduce it, we develop three additional treatments. In our model, correspondence bias is driven by underestimating the correlation between games and actions, which leads to an overestimation of the prosociality of the benign-game player, and an underestimation of the prosociality of the malign-game player. Therefore, we should expect that subjects on average prefer dictator-game givings of the benign-game player to those from a stranger, and prefer dictator givings of a stranger to those from a malign-game player. We test this prediction in Treatment 1, in which subjects choose whether to obtain the dictator-game givings of the person they played either the benign or malign game with or those from a randomly chosen stranger. We find that subjects are willing to pay more for dictator-game givings of a benign-game player compared to a stranger, and are willing to pay more for givings of a stranger compared to a malign-game player. Neither of the two results is consistent with a rational Bayesian model, and they jointly suggest that subjects simultaneously over-estimate the signal value of cooperation in the benign-game and under-estimate the signal value of cooperation in the malign-game.

In Treatment 3, we test whether we can reduce correspondence bias through making people better understand the correlation between game structures and actions. In this treatment, subjects play each

of the games with two different players in Stage 2. In Stage 3, they then choose whether to obtain the dictator givings of their Stage 2 benign- or malign-game co-player. The idea is that subjects can better understand that actions are game-contingent through experiencing both games themselves. Because their own actions will likely be different in the two games, it becomes more clear to them that they should take game structures into account when inferring from actions. Consistent with such an effect, we find that the benign premium in Treatment 3 is smaller than that in Treatment 2, although it is still significantly greater than 0 (at the 1% level), suggesting that experiencing both games is not enough to eliminate the bias.

With Treatment 4, we investigate the effect on reducing the bias of directly showing subjects the correlation between game structure and actions by providing counterfactual information. The setup of Treatment 4 is the same as in Treatment 3, with the exception that subjects are also informed of their benign-game player's action in the malign game and their malign-game player's action in the benign game. In this treatment, as subjects know both players' actions in both games, they should be even more aware of the game-contingent nature of play, which should further reduce the bias. Supporting this prediction, we find that providing counterfactual information reduces the benign premium to 2 cents, which is not significantly different from 0 and is significantly smaller than that in treatments 2 and 3.

The paper proceeds as follows: Section 2 reviews literature in both economics and psychology. Section 3 proposes a conceptual framework for understanding correspondence bias. Section 4 introduces the experimental design and the predictions it tests. Section 5 presents results. Section 6 discusses economic implications of correspondence bias, and Section 7 concludes and discusses policy implications.

2 Literature Review

In this section, we first review related literature in economics, including studies on misattribution and belief updating. We then summarize the extended literature on correspondence bias in psychology, followed by a discussion on how our study differs from the previous studies in psychology and contributes to that literature.

The research that this study is most closely related to in economics is Haggag et al.'s (2019a) investigation of "Attribution Bias in Consumer Choice." In their study, people underweight the impact

of a transitory state, such as hunger, on the utility of consuming a good, and misattribute it to the enduring characteristic of the good. In a related study, Haggag et al. (2019b) find that college students misattribute fatigue generated by being assigned to an early morning section of a course (or back-to-back courses prior to a course) to disinterest in the subject, leading them to subsequently be less likely to major in that subject. The current research builds on their contribution by showing that attribution bias exists not only when it comes to evaluating consumption experiences (or topics of study), but also in evaluating people. Agents in Haggag et al.'s (2019a) model do not fully appreciate the fact that their preferences are state-dependent; similarly, agents in our work fail to fully recognize that actions of other people are game- or incentive-dependent.

Graeber (2020) studied a related but slightly different problem. Subjects in his experiment over-attribute a signal to a payoff-relevant random variable when the realization of the signal is jointly determined by this random variable and another unobserved random variable that is payoff-irrelevant. In contrast, in our environment subjects need to make an inference about a random variable (type of player) from a signal (action) which is jointly determined by this random variable and a known factor (game incentive structure). While the problems are conceptually similar, our environment is less challenging in the sense that subjects do not need to figure out the joint distribution of two random variables. Demonstrating correspondence bias, we show that subjects make systematic mistakes in inference even when the other factor that jointly determines the signal with the random variable of interest is known and salient to them. Complementing Graeber (2020)'s approach, we keep salience constant in our treatments and vary subjects' understanding of the correlation between actions and game structures. We show that the bias in inference can be reduced or even eliminated by making subjects better understand the correlation.

The current research is also related to prior research showing that people respond more strongly to games that they actually play as opposed to those that they observe (Simonsohn et al., 2008), and that people are more likely to make mistakes when inferring information from hypothetical events than from realized events (Esponda and Vespa, 2014, 2019; Martínez-Marquina et al., 2019; Ngangoué and Weizsäcker, 2021). The reduction of correspondence bias in Treatment 3 (observed vs experienced information) and Treatment 4 (providing realized counterfactual information) indicates that people make better decisions when less hypothetical thinking is involved. The same results, showing the importance of personal experience, are also consistent with one of the effects reported in Haggag et

al. (2019a): their finding that past experiences with a good attenuates the attribution bias.

We also contribute to the literature on people's belief updating relative to Bayesian updating. Previous evidence suggests that people generally infer less from evidence than Bayes' Theorem predicts (Phillips and Edwards, 1966; Edwards, 1968; Möbius et al., 2014; Ambuehl and Li, 2018). However, as pointed out by Kahneman, this finding is in contrast to the everyday experience that people often jump to conclusions based on little information. We provide another reason, in addition to the Law of Small Numbers (Kahneman and Tversky, 1972) and base-rate neglect (Kahneman and Tversky, 1973), for why people may draw overly extreme conclusions from small samples.[2] In our case, people jump from observations of others' actions in narrow contexts to conclusions about those people's underlying qualities without paying sufficient attention to the transient incentives they are facing. More interestingly, our results indicate that the same people can both under-infer and over-infer depending on the signals they receive. Even though subjects behave in a way consistent with over-inference when their partners choose to cooperate in the benign game or defect in the malign game, they also tend to infer too little when their partners choose to cooperate in the malign game.

Correspondence bias, also known as the "fundamental attribution bias," has been intensively studied by psychologists since the 1960s (Jones and Harris, 1967; Ross, 1977; Gilbert and Malone, 1995; Gawronski, 2004). In the most common "attitude attribution paradigm" developed by Jones and Harris (1967), subjects read an essay arguing in favor of or against an issue (e.g. Fidel Castro's regime), are informed that the speakers' positions are randomly assigned, and are asked to rate the true attitudes of the speaker towards that issue. The repeatedly-replicated finding is that, despite being informed about the random assignment to positions, subjects still rate the writer who argues in favor of the issue as more supportive of it than the writer who argues against it.[3]

Our work most significantly differs from the past research in psychology in its robustness to

[2]For a discussion of over-inference, see Benjamin (2019).

[3]Ross (1977) developed another popular paradigm for assessing correspondence bias, the "quiz game". In this paradigm, subjects are randomly assigned to be a questioner, a contestant, or an observer in a group of 3 players. The questioner is instructed to compose 10 challenging general knowledge questions and the contestant is asked to answer the 10 questions. Lastly subjects rate the general knowledge of all players in their own group after observing the performance of the contestant.

Bayesian updating. What is interpreted as correspondence bias in this "attitude attribution paradigm" can also be explained by Bayesian updating (Walker et al., 2015). For example in the essay study, even though the title of the essay is pre-determined by the instructor in the no-choice condition, it is hard to imagine students comply uniformly. They may refuse to write an essay favoring positions contradicting their true attitudes.[4] Or at least they should be able to express some of their personal opinions with the given title. Then a Bayesian who saw an essay endorsed a position clearly would also rate the author as more supportive of it than someone who wrote an essay clearly against it. In other words, both the environment and the choice of the individual in that environment are randomly assigned in the "attitude attribution paradigm." The equivalent comparison in our framework would be to let subjects rate the prosociality of a benign-game player who chose to cooperate in the benign game and a malign-game player who chose to defect in the malign game. Instead, we compare subjects' WTPs towards benign-game and malign-game players without conditioning on their choices in the games they play. While Bayesian updating predicts that the benign-game player who chose to cooperate is more prosocial than the malign-game player who chose to defect, it also predicts that benign-game players are in expectation as prosocial as malign-game players. Therefore, we contribute to the psychology literature on correspondence bias by showing that Bayesian updating cannot explain the existence of the bias.

Our work also differs from previous studies on correspondence bias in the salience of the situation. In most studies, following Jones and Harris (1967), the behaviors (a written essay or a videotaped speech) are often very salient, but situational constraints are often insignificant and vague. For example, Choi and Nisbett (1998) inform subjects of the situation faced by the target person by "Please write a short essay in favor of (or opposed to) capital punishment regardless of your own attitude. What is important is your writing skill, not your attitude." It is hard to notice that the target person was forced to write in the assigned direction and it is difficult to evaluate how much pressure there was on the target person to comply with the assigned attitude. Our results indicate that correspondence bias is robust to the salience of the situation. In our framework subjects have complete information

[4]For example, in Sherman (1980) less than 70 percent of university students complied with the request to write a counter attitudinal essay.

about the situations the benign and malign-game players face. By playing the two games themselves, subjects are able to understand how the incentives of the games may shape the behaviors of other players, which is missing in most previous studies. Interestingly, our Treatment 3 results suggest that correspondence bias still exists even when subjects are given experience with the different situations that influence people's behavior, information that one might think would equip them to fully understand the power of situational forces.

The current study augments the existing psychology research on correspondence bias in three other ways. First, the "attitude attribution paradigm" also suffers from the potential confound that subjects may believe that the randomly assigned positions can potentially shape the speakers' attitudes. As we discussed, our design rules this out. Second, in an environment that closely mimics real-life interpersonal interactions, our design clearly shows that correspondence bias not only alters people's assessment of others' attitudes towards an object but also affects people's incentivized choices in economic games. We also show that this bias is welfare-reducing. Third, we empirically establish that counterfactual information can be used to reduce or even eliminate correspondence bias. While previous studies on debiasing correspondence bias focused on general training interventions (Morewedge et al., 2015), we show that a consideration of the causes of correspondence bias can also provide insights into how to debias it.

The most common explanation that psychologists offer for correspondence bias is that, when attempting to make sense of a person's behavior, the characteristics of the person are typically more "salient" than their situation, resulting in an automatic attribution to the former, and an insufficient situational correction (Gilbert, 1989; Gilbert and Malone, 1995; Gawronski, 2004). We formulate the bias in a different way. We are less focused on the salience of other people's characteristics, but more on assessments of their stability. In our formulation in the following section of the paper, it is people's failure to fully account for the incentive-contingent nature of others' actions that leads them to under-attribute actions to incentives. As we show in treatments 3 and 4, when people understand better the correlation between others' actions and the incentive structures through either experience or counterfactual information, correspondence bias decreases or even disappears. This result cannot be explained by the salience-based theory as the salience of situation versus disposition does not change in Treatment 3 or 4.

3 Conceptual Framework

In this section, we build a simple descriptive model of correspondence bias. In our framework, the individual does not fully take into account the fact that other people's actions depend on the incentives they face (or the game they play); they are aware of the distributions of others' actions, but underestimate the correlation between actions and the game structure when they try to interpret those actions.

Basic Setup of the Model

Consider two games $\tau \in \{b, m\}$, the *benign game* b and the *malign game* m. Both b and m are symmetric two-player complete information games. There are two types of agents $t \in \{G, B\}$, the Good type G and the Bad type B. Let the probability of being the Good type be $p_0 \in (0, 1)$. There are two actions to take in both b and m: $a \in \{C, D\}$. In the benign game b, both the Good type and the Bad type choose C in equilibrium; in the malign game m, the Good type chooses C and the Bad type chooses D in equilibrium.[5] Half of the players are assigned to play the benign game, and the other half are assigned to play the malign game. Let player i be a player who is assigned to the benign game and player j be a player who is assigned to the malign game.

After observing player i's action in the benign game and player j's action in the malign game, a risk-neutral player k chooses between i and j to play a follow-up game. k's payoff in the follow-up game is defined by the type of the partner of her choosing. The Good type is preferred by every player. Specifically, we standardize the payoff of having a type B player as the follow-up game partner to 0 and having a type G player to 1.[6] Player k's expected payoff for choosing player $l \in \{i, j\}$ play the

[5] As the focus of this paper is on belief updating, we abstract away from the details of the two games. Another way to look at how the two types of players act in the two games is that we *define* players who choose C in both games as the Good type, and players who choose C in the benign game and D in the malign game as the Bad type. In our experiment, the benign game is the harmony game in which it is a dominant strategy to cooperate, and the malign game is the prisoner's dilemma game. Subjects are not aware of the future stages when they play the two games. Thus they have no incentives to hide their types when playing the games.

[6] One way to understand this assumption is to view the Bad type as the selfish type and the Good type as a behavioral type. The bad type only cares about his own welfare. Thus he would find it to be a dominant strategy to defect in the

follow-up game with is

$$U_{kl} = p(t_l = G), \tag{1}$$

where t_l is the type of player l and $p(t_l = G)$ is the true probability of player l being the Good type.[7] If U_{ki} is larger than U_{kj}, it means the expected payoff of choosing player i is higher than that of choosing player j.

The Bayesian Benchmark

We first describe how a rational Bayesian behaves in this environment. What this Bayesian needs to do is to interpret the action of player i in the benign game and the action of player j in the malign game, and form posterior beliefs about i being a Good type and j being a Good type. Then she chooses the player who is more likely to be the Good type to play the follow-up game with. It is a relatively simple decision for a Bayesian. Intuitively, benign-game player i and malign-game player j are equally likely to be the Good type *ex ante* because which game a player is assigned to is determined by chance. As the expected posterior is equal to the prior, due to the Martingale property of Bayesian updating, i and j are equally likely to be the Good type in expectation *ex post*. We summarize this intuition in the following lemma.

Lemma 1. *A risk-neutral Bayesian is in expectation indifferent between a benign-game player and a malign-game player to play the follow-up game with.*

Proof. To prove this lemma, we look at k's posterior beliefs about benign-game player i's type and malign-game player j's type. Define $\pi(\cdot)$ as an individual's (potentially biased) belief. As both types

prisoner's dilemma game. In contrast, the Good type chooses to cooperate in the prisoner's dilemma game in light of an other-regarding preference or a preference for efficiency. The follow-up game in our experiment is the dictator game, and the Good type is supposed to transfer more to the recipient. We find support for this assumption. Players who chose to defect in the prisoner's dilemma game transferred 60.17 in the dictator game, while players who chose to cooperate transferred 76.59, a difference that is significant at the 1% level. In contrast, players who chose to cooperate in the harmony game did not transfer more in the dictator game than players who chose to defect in the harmony game (67.11 vs 62.9, p=0.357).

[7] By formulating the expected payoff in this way, we assume that the decision maker is risk-neutral. However, our main results remain unchanged by assuming risk aversion.

choose C in the benign game, the posterior is equal to the prior, $\pi(t_i = G \mid a_i^b = C) = p_0$, where a_i^b is player i's action in game b. In the malign game, player j's type is perfectly revealed. If she chooses C, then she is surely the good type, $\pi(t_j = G \mid a_j^m = C) = 1$; if she chooses D, then she is surely the bad type, $\pi(t_j = G \mid a_j^m = D) = 0$. The expected posterior $E[\pi \mid \tau = m] = p_0\pi(t_j = G \mid a_j^m = C) + (1 - p_0)\pi(t_j = G \mid a_j^m = D) = p_0$. As both expected posteriors are equal to the prior, p_0, they are also equal to each other. Therefore, the expected payoff of choosing a benign-game player and choosing a malign-game player is the same. □

Although our theoretical analysis assumes that the decision-maker is risk-neutral, it is worth considering how risk-aversion would affect the behavior of a Bayesian in our environment. Lemma 1 describes how a risk-neutral Bayesian chooses between a benign-game player and a malign-game player. A somewhat surprising result is that a risk averse Bayesian – who is not subject to correspondence bias – should, in expectation, prefer a malign-game player to a benign-game player. The intuition is straightforward. A player's behavior in the malign game better reveals their type than does a player's behavior in the benign game. This is true empirically, in our experiment: dictator-game giving is better predicted by a player's behavior in the malign game than their behavior in the benign game. In our model we make the assumption that play in the malign game perfectly reveals the type of the player: the Good type chooses action C, and the Bad type chooses action D in equilibrium. After observing a player's behavior in the malign game, therefore, there is no *ex post* uncertainty regarding the type of the player. In contrast, the benign game reveals no information on the type of its players because both types choose action C in equilibrium. As the expected payoff of choosing the benign-game player and the malign-game player is the same, a person who is risk averse but not subject to correspondence bias should, on average, choose the malign-game player.[8]

Correspondence Bias and Its Implications

While a Bayesian is in expectation indifferent between a benign-game player and a malign-game player, the same may not be true for a correspondence-biased individual. In this paper, we define cor-

[8]This is, of course, the opposite of the pattern of behavior that we observe, which suggests that risk aversion cannot explain our results and, if anything, leads to an underestimation of the magnitude of the correspondence bias.

respondence bias as the failure to fully account for the incentive structure of a game when interpreting a player's actions in the game. Borrowing from Eyster and Rabin's (2005) *cursed equilibrium*, we define an agent as *correspondence-biased* if her posterior belief about player l's type given her action a_l^τ in game τ corresponds to:

$$\pi(t_l = G \mid a_l^\tau) = \chi[p(b \mid a)p(t_l = G \mid a_l^b) + p(m \mid a)p(t_l = G \mid a_l^m)] + (1-\chi)(p(t_l = G \mid a_l^\tau)), \quad (2)$$

where $p(\tau \mid a)$ is the probability of the game being τ given action a, and $\chi \in (0,1]$ is the probability that the individual only recognizes the action of her opponent but ignores the incentive/game structure she faces. When she is unable to recognize the incentives her opponent faces, she replaces the actual probability $p(t_l = G \mid a_l^\tau)$ of her opponent being the Good type given action a in game τ with the average posterior of her opponent being the Good type given action a across the two games, $p(b \mid a)p(t_l = G \mid a_l^b) + p(m \mid a)p(t_l = G \mid a_l^m)$. If $\chi = 0$ instead, then the biased individual's posterior is the same with a Bayesian.

To see how a correspondence-biased individual acts differently from a Bayesian, we look at her belief updating when facing a benign-game player and when facing a malign-game player. Intuitively, a correspondence-biased individual tends to over-react to action C in the benign game and under-react to action C in the malign game. While action C in the benign game conveys no information about a player's type, action C in the malign game is a strong signal for the Good type. When the correspondence-biased individual is unsure in which game an action is taken, there are chances that action C in the benign game is interpreted as action C in the benign game, and *vice versa*. This leads to an overestimation of the probability of a benign-game player being the Good type and an underestimation of the probability of a malign-game player being the Good type. We formalize this intuition in the following lemma.

Lemma 2. *A correspondence-biased individual prefers a benign-game player to a stranger to play the follow-up game with, and in expectation prefers a stranger to a malign-game player to play the follow-up game with.*

Proof. As the chance of being the Good type is p_0 for the stranger, we need to compare the expected posterior beliefs of the correspondence-biased individual with p_0 to prove this lemma.

For the benign-game player i, the correspondence-biased individual's posterior belief $\pi(t_i = G \mid a_i^b)$ is larger than p_0. To see this, $\pi(t_i = G \mid a_i^b = C)$ in this case equals to $\chi(p(b \mid C)p(t_i = G \mid a_i^b = C) + p(m \mid C)p(t_l = G \mid a_i^m = C)) + (1-\chi)(p(t_i = G \mid a_i^b = C))$. As $p(t_i = G \mid a_i^m = C) > p(t_i = G \mid a_i^b = C)$, it follows that $\pi(t_i = G \mid a_i^b = C)$ is larger than p_0.

For the malign-game player j, we can derive that $\pi(t_j = G \mid a_j^m = C) < p(t_j = G \mid a_j^m = C) = 1$ by applying the same logic in the above paragraph. At the same time, $\pi(t_j = G \mid a_j^m = D) = p(t_j = G \mid a_j^m = D) = 0$ as action D can only appear in the malign game. This indicates that $E[p(t_j = G \mid \tau = m)] > E[\pi(t_j = G \mid \tau = m)]$. As $E[p(t_j = G \mid \tau = m)] = p_0$, $E[\pi(t_j = G \mid \tau = m)]$ is smaller than p_0. □

As a correspondence-biased individual prefers a benign-game player to a stranger, and at the same time prefers a stranger to a malign-game player in expectation, a natural corollary is that a correspondence-biased individual would prefer a benign-game player to a malign-game player in expectation. This gives us the main result of the model.

Proposition 1. *A correspondence-biased individual in expectation prefers a benign-game player to a malign-game player to play the follow-up game with.*

This proposition implies that a correspondence-biased individual is willing to pay a premium for the benign-game player. We call this premium the *benign premium*.

Definition. We define a correspondence-biased agent's *benign premium* as her expected payoff of choosing the benign-game player over the malign game player, namely $E[\pi \mid \tau = b] - E[\pi \mid \tau = m]$.

We utilize the *benign premium* to test for the existence of correspondence bias. While a correspondence-biased individual is willing to pay a positive *benign premium*, a Bayesian is, in expectation, willing to pay 0 for the benign-game player.

4 Design

The experiment has three stages. In the first stage, all subjects make a decision as the dictator in the dictator game. In the second stage, they are randomly matched into groups of 4 to play the benign game and the malign game. The benign game was chosen to encourage players to cooperate with the other player, while the malign game was chosen to motivate selfish behavior. Lastly, they are asked, as the receiver, to choose between receiving the dictator givings of two players from the first stage.

Our model predicts that there exists a *benign premium*: subjects are, on average, willing to pay to be matched with the benign-game player.

First Stage

The experiment was conducted online, and subjects were recruited through Amazon Mechanical Turk (Mturk). Upon arriving at the study website, each subject was instructed to play a dictator game as the dictator. They divided 200 cents between themselves and a random receiver. As in a standard dictator game, the receiver had no influence over the outcome of the game, and both the receiver and the dictator receive 50 cents of endowment prior to the split decision. Subjects were also informed that, although everyone needed to make the decision, only half of those decisions would be implemented later. At this stage, they had no idea of the existence or nature of the future stages of the experiment or of the identity of the potential random receiver. This dictator decision serves as our measure of each subject's prosociality.

Second Stage

In the second stage, subjects were randomly matched into four-player groups. Everyone was randomly assigned a role. There were four roles in each four-player group. We name them A, B, C and D. Then, the participants played the benign and/or the malign games with individuals in their own group. Depending on the treatment, a subject interacted with one or two individuals at this stage. The two games are defined as follows.

The malign game is a two-player one-shot prisoner's dilemma, see the left panel in Table 1. Participants in this game must choose between cooperate (C) and defect (D).[9] It is a dominant strategy to choose D for both players. The Nash equilibrium of this game is (D,D), which leads to the payoffs (30,30). Even though D is the dominant strategy, in previous studies not everyone defected, and those who chose to cooperate were more prosocial, as measured by givings in a dictator game, than those who choose to defect (Cooper et al., 1996; Barreda-Tarrazona et al., 2017). In the language of the

[9]The actions C and D are respectively labeled "Action 1" and "Action 2" in the experiment to ensure a neutral presentation.

model, subjects who cooperate in the malign game are *Good* types and subjects who defect in the malign game are *Bad* types.

The benign game is the Harmony game as in Dal Bó et al. (2018) (right panel in Table 1). Participants also choose between cooperate (C) and defect (D). However, (C,C) is the dominant strategy equilibrium and Pareto dominates all other strategy profiles in this game. It is easy for the subjects to figure out that they should choose C, and most do so.

To ease understanding, we illustrate the rest of the experimental design in terms of Treatment 2, which we believe to be closest to situations encountered in daily life. We discuss the differences between the treatments at the end of this section. In Treatment 2, subjects play one game, and observe outcomes of the other game. Specifically, players A and B play the benign game, and C and D play the malign game. Even though they only play one game, we also give them the instructions, including payoffs, of the other game so that they can still understand the incentives of the game they are not playing.

After Stage 2 actions were taken, subjects entered the information provision page. On this page, they learned about their payoffs and the action of their opponent in the game they had just played. We also displayed the payoff table of the two games again to minimize confusion and to aid in recall. In addition, in Treatment 2, subjects were informed of the action of one of the two players in the game they did not play. To be more precise, A (B) in the benign game also learned whether D (C) in the malign game chose to cooperate or not. Similarly, C (D) was also informed of the action of B (A) in the benign game.

Subjects were forced to stay on the information provision page for at least 120 seconds to make sure that they had the time to understand the game structure and make inferences about the types of a benign-game player and a malign-game player based on their actions.

Third Stage

In the third stage, every subject chose whether to receive the dictator transfer from the first stage either from a malign-game player and a benign-game player. In the second experimental treatment, one of them was the player they had played with, and the other was the player whose play they only learned about. The two candidates are those two whose actions in Stage 2 were shown to the subject in the information provision phase. For example, A played the benign game with B and observed

D's action in the malign game. Then in Stage 3, A chose between B and D's dictator transfers in the first-stage dictator game. Therefore, the only source of information that the subject had to go on when choosing between the two candidates was the action that each had taken in the game they played in Stage 2 (as well as the payoffs for these games).

After reading the instructions for Stage 3, and before making any decisions, subjects were asked three comprehension questions. Only those who answer all three questions correctly could proceed to make their choices in this stage; those who answered at least one question incorrectly were required to re-do all three questions until they answered them all correctly.[10]

After a subject made the choice between the two candidates, we used a multiple price list to elicit her WTP to be matched with the candidate of her choosing. The list shown to A if she chooses B over D in the first choice is displayed in Table 2 as an example. In total, subjects made 10 choices, excluding the first one. In each choice, there were two options. D+(x cents) means if in this choice A chooses D, and this is the choice selected at random to count, then she will then get an extra reward of x cents. But if she chooses B, there is no extra reward. The point where A switches from option 1 to option 2 defines A's WTP to get B. One of these 11 choices (including the one between B and D with no extra rewards) was randomly selected as the *choice-that-counts*, and the instructions made this clear. We then implement one of the four *choice-that-counts* with a designated matching protocol.

Our matching protocol in Stage 3 was designed to eliminate a potential confound of correspondence bias. If we had designed the experiment differently, one way to reciprocate cooperation by the benign-game player could have been to choose her as the dictator in Stage 3, as dictators are expected to earn more than receivers. This was not, in fact, an issue because, if a player was designated to be the dictator by another player's choice, the unchosen dictator was assigned to be the dictator for another player who did not get to choose. So, a player's choice of dictators did not determine who became a dictator, which was a matter of chance. More specifically, the protocol can be divided into 4 steps. In the first step, we randomly chose 1 player from the 4. Let us name her the *chosen player*

[10] As one cannot proceed to the decision stage of Stage 3 without answering all 3 questions correctly, some subjects dropped out in this stage. Out of 1,008 subjects who signed up for the experiment, 151 of them finished Stage 2 but dropped out in Stage 3. As Stage 3 is not interactive, the dropout of those subjects has no impact on the use of data from others in the same group. Moreover, there is no significant difference in attrition rate across treatments.

and assume it was A. Second, both the *chosen player*'s (A) benign- and malign-game players (B and D) got the dictator role. Therefore who played as the dictator and who played as the receiver in the first stage dictator game was determined at the second step. The next steps only affected the matching between the two dictators and the two receivers. In the third step, we implemented the *chosen player*'s (A) *choice-that-counts*. Suppose A chose B's dictator offerings in that choice, then A and B were matched with B as the dictator. Fourth, the player who was not picked by the *chosen player* A, D in our example, was matched with the remaining player C. D's dictator transfer decision was carried out, and C received D's offerings as the receiver. The fact that D was the dictator but not C was determined in step 2. Therefore, who got dictator roles was completely determined by whom was randomly selected to be the *chosen player* in the matching protocol, choosing a player as the dictator in Stage 3 did not raise her chances of being the dictator in the dictator game.

At the end of Stage 3, the dictator's decisions made in Stage 1 were then carried out. For example, suppose B chose to give x cents to the random receiver in the first stage, and if A and B were matched, with B being the dictator, then A received x cents and B received $(200-x)$ cents. Putting the dictator decision ahead of the second-stage games eliminated the possible confound, if the dictator decision had followed the second stage games, that people's experience in a game can influence their prosociality. By the third stage the dictators had already made their decisions about how much to transfer in the first stage, so what happened at the second stage could not have an impact on their behavior. Even if the benign-game player became a nicer person after playing the game, her choice in the first stage remained the same.

Treatments

There are four treatments in the experiment, and they only differ in the second stage. What differentiates them from each other is how many games each subject plays and how much information they are given.

In Treatment 1 (as indicated in Figure 1), each player only plays one game, either the benign or the malign game, and is not aware of the existence of the other game. In the third stage, subjects are asked to choose between receiving the dictator-game givings of the person they play this one game with, or those from a random participant in the study.

In Treatment 2, as already described, each player again only plays one game. However, in this

condition, in addition to the outcomes and the action of the player in the game he/she plays, the subject also observes the action of one other player who plays the other game, as well as receiving full information about the game itself. Then, the focal subject chooses whether to receive the dictator-game givings of the player who he/she actually played with, or the player who he/she only received information about.

Treatment 3 is the same as Treatment 2, except that subjects actually play both games (the benign game and the malign game) with different other subjects in their group. In the information stage, they learn the actions of both of the players they play with. The difference between Treatment 3 and Treatment 2, therefore, is that in Treatment 3 all information is gathered through "experience" instead of partly through "observation" as in Treatment 2.

Treatment 4 is the same as Treatment 3 except that subjects are also informed of the behaviors of their benign-game player in the malign game and the behaviors of their malign-game player in the benign game. For example, if A plays the benign game with B, she also learns how B behaved playing the malign game with D.

Predictions

The four-treatment design helps us investigate the mechanisms behind correspondence bias and the potential ways to reduce or even eliminate it.

Prediction 1. There exists a benign premium in Treatment 2, that is, the average WTP towards the benign-game player's dictator-game givings is larger than that towards the malign-game player.

Treatment 2 is our baseline treatment, and we can test the existence of correspondence bias by looking at the benign premium in this treatment.[11]

Prediction 2. In Treatment 1, when choosing between a benign-game player and a random stranger,

[11] We choose Treatment 2 as our baseline for two reasons. First, in daily life, people often draw inferences about others' characteristics based on their personal experience with those people, but with only second-hand knowledge of those people's behavior in other environments. Second, Treatment 2 is directly comparable to Treatment 3 and 4, as in all three of these treatments subjects chose, in Stage 3, between a benign-game player and a malign-game player's dictator-game givings. In Treatment 1, in contrast, they chose between a benign or malign player and a stranger's givings.

subjects are on average willing to pay more to receive the dictator-game givings of the benign-game player; when choosing between a malign-game player and a stranger, they are on average willing to pay more to receive the dictator-game givings of the stranger.

Treatment 1 aims to decompose the benign premium. As no information is provided on the stranger, the chance of her being the Good type is equal to the prior, p_0. Thus, Treatment 1 helps us separate the benign premium into two parts: underestimation of the chance of the malign-game player being the Good type and overestimation of the chance of the benign-game player being the Good type. While Bayesian inference predicts that willingness to pay to receive the benign-game and malign-game players' dictator-game givings should be the same as the willingness to pay to receive the stranger's givings, as a result of correspondence bias, we predict that agents will prefer the benign-game player's givings to the stranger's, and prefer the stranger's givings to the malign-game player's.

Prediction 3. The benign premium is smaller in Treatment 3 than in Treatment 2.

Treatment 3 is set to test whether misunderstanding of the correlation between behaviors and strategic motives is a cause of correspondence bias. As participants play both games in this treatment and likely make different choices in the two games, they have a better understanding of how incentives in the two games influence players' actions. We expect the benign premium to shrink in Treatment 3 compared to Treatment 2.

Prediction 4. The benign premium is smaller in Treatment 4 than in Treatment 3.

In Treatment 4, we test whether providing counterfactual information reduces correspondence bias. In treatments 2 and 3, participants are not able to know how the benign-game players perform in the malign game, and vise versa. However, in Treatment 4, such information is available, and subjects can clearly see how others' actions change according to the incentives they face. If correspondence bias is caused by failing to fully account for the impact of the incentives on actions, then enabling people to compare other players' behaviors in different games with different incentives should reduce the bias significantly.

Implementation

The experiment was conducted on Amazon Mechanical Turk between October 12, 2018 and December 7, 2018. As our experiment is rather complicated, we only recruited subjects who had at least

a two-year associate degree. We also restricted participation to residents of the United States who had completed at least 100 tasks prior to our study and had an approval rating of at least 95%. We advertised the experiment as a 20-minute academic decision-making study with an average payment of 2.5 dollars. On average, the experiment lasted 20.1 minutes and subjects earned 2.77 dollars.

Overall, we recruited 817 subjects in our online experiment, 121 in Treatment 1, 246 in Treatment 2, 223 in Treatment 3, and 227 in Treatment 4.[12] We randomly assigned fewer subjects to Treatment 1 based on a power calculation. We needed more subjects in the other 3 treatments because we tested whether the benign premium is significantly different between every two treatments, whereas in Treatment 1, we only need to test whether the average WTP is significantly different from 0 or not.

Table 3 shows summary statistics both in aggregate and across treatment conditions. All of the non-outcome behaviors and demographics are balanced. On average, subjects shared 67 cents in the dictator game. 95.2% of subjects chose to cooperate in the benign game and 38.9% defected in the malign game. A natural concern is that subjects may behave differently in Treatment 2 and in treatments 3 and 4, because the number of games they play is different. Reassuringly, the cooperation rate in the malign game in Treatment 2 is not significantly different from the average cooperation rate in treatments 3 and 4 (p-value=0.424). We collected subjects' demographic information in a voluntary follow-up survey. 735 out of 817 subjects (90%) completed the survey, and there is no significant difference in the take-up rates across treatments. Survey respondents have an average age of 38, 57% are female, and 80% have jobs (either employed or self-employed).

5 Results

The objective of this study is to examine whether, when people make inferences about others based on their behaviors, they over-attribute behaviors to others' characteristics and underestimate the impact of incentives on behaviors. To do so, we look at how an individual's randomly assigned game, which is orthogonal to her characteristics, affects other people's perceptions of her. We first confirm that the game a subject is assigned to play is indeed orthogonal to her prosociality, which is measured

[12] We received a total of 857 responses, but dropped 40 subjects (4.67%) who exhibited multiple switching points in the multiple price-list questions at the third stage.

by her dictator givings in Stage 1. Figure 2 (right-hand panel) illustrates that subjects who play the benign game transfer an average of 66.90 cents, which is, as would be expected if randomization was successful, almost identical to the average dictator givings from malign-game players (66.56 cents; p-value=0.89).

Then as a manipulation check, we look at whether the two games induce different behaviors (Table 4). While almost everyone (95.2% of subjects) chooses to cooperate in the benign game, the frequency of cooperation is much lower in the malign game (38.9%), so the game structure does indeed affect subjects' choices. The choices in the malign game are also informative for identifying types of subjects. Figure 2 shows that subjects who choose to cooperate in the malign game transfer 77 cents in the first stage, while subjects who choose to defect only transfer 60 cents, a statistically significant difference (p-value$<$0.01, rank-sum test). In contrast, though those who cooperate in the benign game do, on average, contribute more (67 cents) than those who do not cooperate (63 cents), the difference does not approach significance (p-value=0.36).

Result 1. *Correspondence bias exists in the baseline treatment when subjects experience the action of one player and observe the action of another player. The existence of the bias leads to a clear welfare loss.*

Turning to the main results of the paper, we first look at the existence of correspondence bias in the baseline treatment, Treatment 2. A rational Bayesian model predicts that subjects will be, in expectation, indifferent between receiving the dictator offerings from either the benign-game player or the malign-game player. However, supporting the first prediction of our model, there is a positive benign premium: subjects are willing to pay, on average to receive the dictator game offerings from the benign-game player rather than those from the malign-game player. Using the multiple price list, we define the willingness-to-pay (WTP) for the benign-game player as the switch point between option 1 and option 2 in Table 2. We further code it as positive if a subject chooses the benign-game player in the first choice, and negative otherwise. Since the multiple price list can only elicit intervals of WTP, we use the mid-point of the interval as the WTP for the benign-game player.[13] For example, if

[13]The results are robust if we use the lower or upper bound of the interval as the WTP for the benign-game player (Figure 5).

subject A chooses B's (the benign-game player) transfers over D's (the malign-game player) transfers plus 10 cents bonus, and switches to D's transfers plus 20 cents when choosing between it and B's transfers, then A's WTP for the benign-game player is coded as 15 cents.

As shown in Figure 3 and Table 5, the average WTP for the benign-game player's dictator givings is 11.67 cents higher than that for the malign-game player's givings in Treatment 2, which is significantly larger than 0 at the 1% level. The Bayesian model is rejected. One way to interpret this result is that subjects believe that the benign-game player on average transferred 11.67 cents more in Stage 1 than the malign-game player. To put those numbers into perspective, one can compare them with the maximum plausible benign premium of 100 cents. A completely selfish individual transfers 0 in Stage 1, while an altruistic individual who weights others' utility exactly as much as her own transfers 100 cents in Stage 1. Therefore, although larger values are possible (up to 200 cents), the largest plausible difference between the two players' transfers is 100 cents.

The benign premium can also be interpreted as a measure of the welfare loss caused by correspondence bias. To see this, consider the case when the expected dictator givings of the malign-game player are higher than that of the benign-game player from a Bayesian's perspective but the difference between the two is smaller than the benign premium. While a risk-neutral Bayesian would choose the malign-game player, a risk-neutral correspondence-biased agent would still choose the benign-game player, leading to an expected welfare loss. The larger the benign premium, the more likely a correspondence-biased agent would forfeit a gain from choosing the malign-game player's dictator givings.

Given that, at the aggregate level subjects are correspondence-biased, a natural next question is how many subjects are correspondence-biased. This question is hard to answer when the malign-game player chooses to defect. Both the Bayesian model and our model predict that in this situation subjects *should* choose the benign-game player, and the only difference is that our model predicts a larger WTP towards the benign-game player. However, the case when the malign-game player chooses to cooperate is clear-cut. While a Bayesian subject should choose the malign-game player regardless of her prior, our model predicts that a fully correspondence-biased subject is indifferent between the two players and may choose the benign-game player. Consistent with this prediction, our data show that 52% of subjects choose the benign-game player over the malign-game player when the latter choose to cooperate in Treatment 2 (Panel A of Table 6).

Result 2. *Evidence suggests that correspondence bias is caused by both an overestimation of the prosociality of the benign-game player and an underestimation of the prosociality of the malign-game player.*

In Treatment 1, subjects only play one game, and are asked to choose between receiving the dictator-game givings of the person they play this one game with, and those from a random participant. As predicted by the model, a Bayesian subject should be indifferent between her partner and a stranger in expectation regardless of which game she is assigned to play. However, the game an individual plays does have an impact on her WTP towards her partner.

Treatment 1 is more comparable to previous studies in psychology on correspondence bias. We randomly assigned subjects to interact with someone in a benign environment (corresponding to the "against an opinion" condition in the psychology literature) or a malign environment (corresponding to the "in favor of an opinion" condition), and we test whether this randomly assigned environment had an impact on a subject's evaluation of their partner or not (corresponding to asking subjects to rate the attitudes of the speaker towards that opinion). Our results show that the orthogonal environment has a strong effect on a subject's WTP towards her partner. When the game played together is the benign game, the average WTP for partners over the strangers is 12.62 cents; when it is the malign game, the average WTP for partners is -7.24 cents, meaning subjects are willing to pay to receive the dictator givings from random strangers, rather than from their partners. The two WTPs are significantly different from each other (p-value$<$0.01, Wilcoxon rank-sum test), which serves as another piece of evidence of correspondence bias.

Treatment 1 also serves as a test of the mechanisms behind correspondence bias. If the bias is caused by people's failure to fully account for the degree to which incentives affect actions, then we would predict a preference for the benign-game player to the stranger and a preference for the stranger rather than the malign-game player. The results are consistent with this prediction. As shown above, the average WTP for the benign-game player is positive and is significantly different from 0, with a p-value of 0.025. Meanwhile, the average WTP for the malign-game player is negative (p-value=0.155). The negative WTP for the malign-game player is unlikely to be a mistake, as subjects do respond to the malign-game player's actions. When the malign-game player chooses to cooperate, the average WTP towards her is 11.67 cents; when the malign-game player chooses to defect, the average WTP is -20.59 cents.

Result 3. *Direct experience with both games reduces correspondence bias, but by itself is not sufficient to eliminate the bias.*

So far these results demonstrate the existence of correspondence bias: subjects tend to believe that someone who they are randomly assigned to play a benign game with is more prosocial than someone who they are randomly assigned to play a malign game with. The next question is whether we can alleviate this bias. By comparing Treatment 2 with Treatment 3, we can see the effect of letting subjects experience both regimes, so as to better understand the correlation between strategic motives and actions. The only difference between the two treatments is that subjects only play one game but observe the other one in Treatment 2, while in Treatment 3 they play both. The average benign premium decreases from 11.67 cents in Treatment 2 to 7.78 cents in Treatment 3, with a p-value of 0.263. This shows that experience alone is not sufficient to eliminate correspondence bias. The benign premium in Treatment 3 is still significantly larger than 0 (p-value=0.003, t-test).

The reduction in the benign premiums from Treatment 2 to Treatment 3 is mainly driven by the reduction in WTP for the benign-game player of subjects whose malign-game player chooses to defect. As shown in Figure 4, when the malign-game player chooses D, the average WTP for the benign-game player decreases from 20.68 cents to 15.16 cents (p-value=0.159). Meanwhile, the average WTP for the benign-game player only decreases from -0.05 cents to -2.16 cents when the malign-game player chooses to cooperate. These results suggest that experience is better at reducing the overestimation of the niceness of the benign-game player. It has little effect on reducing the underestimation of the niceness of the malign-game player.

Result 4. *Providing counterfactual information in addition to letting subjects experience both games eliminates correspondence bias. The result is mainly driven by a reduction in overestimation of the niceness of the benign-game player.*

By comparing treatments 3 and 4, we can study the effect of informing the subjects of "counterfactuals." When, in Treatment 4, we not only let subjects learn the behaviors of two players by playing games with them, as in Treatment 3, but also inform them of the behaviors of the two players in the game they did not play together, the benign premium further decreases to 2.14 cents, which is not significantly different from zero (p=0.407). The difference in the benign premium between Treatment 3 and Treatment 4 is significant at the 10% level (p-value=0.095), suggesting that providing coun-

terfactuals can alleviate correspondence bias. The difference between Treatment 2 and Treatment 4 is significant at the 1% level (p-value=0.007), which indicates that experience plus counterfactual information can jointly eliminate the bias.

The reduction in the benign premiums from Treatment 3 to Treatment 4 is mainly driven by the reduction in the benign premium when the malign-game player chooses to defect (Figure 4 and Table 6). In this situation, the average WTP for the benign-game player decreases from 15.16 cents in Treatment 3 to 6.37 cents in Treatment 4 (p-value=0.103). The average WTP in Treatment 4 (6.37 cents) is very close to the Bayesian level with the correct prior, 6.94 cents.[14] This suggests that the overestimation of the niceness of the benign-game player is almost gone in Treatment 4. At the same time, when the malign-game player chooses to cooperate, the benign premium declines from -2.16 cents in Treatment 3 to -3.65 cents in Treatment 4. Again, it is also closer in Treatment 4 than in Treatment 3 to the Bayesian amount with the correct prior, -9.48 cents.[15]

The finding that providing counterfactuals reduces the correspondence bias helps to explain its robustness in daily life: it is usually impossible to observe the counterfactual behavior of the people we interact with. For example, in a society with low mobility, the rich are born rich and the poor typically remain poor. It is hard to see how the rich would behave if they were poor, and it is hard to observe how the poor would behave if they were rich. Even if some people experienced both cases, others are unlikely to witness how they behave in the two different situations.

Interestingly, even though the benign premium becomes smaller in Treatment 3 and even more so in Treatment 4, the proportion of subjects who are biased remains quite stable. Around 52.81% and 47.31% of subjects in treatments 3 and 4 respectively still choose the benign-game player in the first choice when the malign-game player chooses cooperation, which is inconsistent with the predictions of the Bayesian model but consistent with our model of correspondence bias. One plausible

[14] When subjects are Bayesian with correct priors, the WTP for the benign-game player should be equal to the conditional expected differences in the two players' dictator givings. As Figure 2 illustrates, the difference in the dictator givings from the benign-game player who chooses to cooperate (67.11 cents) and the malign-game player who chooses to defect (60.17 cents) is 6.94 cents.

[15] -9.48 is the difference in the dictator givings between the benign-game player who chooses to cooperate (67.11 cents) and the malign-game player who also chooses to cooperate (76.59 cents).

interpretation is that being correspondence-biased is a relatively stable trait, but that experience and counterfactual information can reduce the magnitude of the bias.

Robustness Checks

One potential concern is our results are driven by the complexity of the design or by subjects' inattention. We use education level to proxy mathematical/computational skills, and test whether people who have fewer years of education show a stronger sign of correspondence bias. For inattention, we use how long subjects stay in each stage as a proxy; people who pay more attention to the study are likely to stay longer in each stage before making their decisions. We present the results in Table 7. In this analysis, we only include observations in treatments 2, 3, and 4, as the definition of the benign premium is slightly different in Treatment 1. When looking at the effect of education on the level of the bias, we continue the analysis with a subsample of subjects who finished the voluntary follow-up survey. As shown in Table 7 column 3, the level of education has no significant impact on the WTP for the benign-game player. The same applies to all the stay-duration variables. In combination, these results suggest that the observed effects are not driven by people with a relatively low level of education or by people who did not pay enough attention to the study.

6 Discussion

In the Discussion section, we first discuss two potential alternative explanations and explain why they cannot explain the entirety of our results. We then present a set of applications of correspondence bias in managerial decision making.

Alternative Explanations

In our baseline treatment, Treatment 2, subjects are willing to pay a benign premium for the benign-game player's dictator-game givings, even though players are randomly assigned to the benign game and the malign game. Subjects in Treatment 2 only play one game, but observe the other game. While we believe this feature closely mimics reality, its asymmetry also opens doors for alternative explanations. In this subsection, we show that results in Treatment 3 and Treatment 4 provide evidence against alternative explanations.

The first alternative explanation to the benign-premium in Treatment 2 is that subjects fail to

understand the strategic situation of the game that they do not play. Thus they interpret defection in the malign game as a bad signal and interpret cooperation in the benign game and the malign game as equally strong good signals, which could also result in a benign premium.

Even though this mechanism may be contributing to some of the effects observed in the experiment, it cannot explain the whole set of our results. We first test whether subjects truly do not understand the game they do not play by examining the behavior of subjects who only play the benign game while observing the malign game. If subjects do not understand the strategic situation of the game they do not play, or pay no attention to it, then we would expect that their willingness to pay for the benign-game player is the same regardless of how the malign-game player behaves. However, Table 6 shows that for subjects who only played the benign game, when the malign-game player chooses to cooperate, their average WTP for the benign-game player is 13.89 cents; while when the malign-game player chooses to defect, the average WTP for the benign-game player increases to 27.88 cents. The two amounts are significantly different from each other (p-value=0.042). This suggests that subjects on average understand how actions in the game they do not play should be interpreted.

Most importantly, if the benign premium in Treatment 2 is entirely driven by people's misunderstanding of the strategic situation of the game they do not play, then we should expect it to disappear in Treatment 3 in which subjects play both games. However, the benign premium still exists in Treatment 3. Subjects are on average willing to pay 7.78 cents for receiving the dictator offerings from the benign-game player, which is significantly greater than 0 at the 1% level. The misunderstanding of the game one does not play explanation cannot explain our results in Treatment 4 either. Subjects play both games in both Treatment 3 and Treatment 4. Thus there is no difference in how many games they play between the two treatments. Nevertheless, we find that providing counterfactual information in Treatment 4 significantly lowers the benign premium from 7.78 in Treatment 3 to 2.14 in Treatment 4.

The second alternative explanation is people may prefer to interact again with a player who they share a good experience with, and may prefer to not interact again with a player who they share a bad experience with. As subjects tend to share a good experience with benign-game players and share a bad experience with malign-game players, a preference for the good-experience player could, in theory, explain why subjects are willing to pay a benign premium.

The three-stage feature of our design is, however, designed to eliminate any impact of a preference

for the good-experience partner. Even if subjects would prefer to interact again with a partner who they share a positive experience with, they cannot realize this preference because they won't interact with their partners again. Their choice in Stage 3 is to choose between the benign-game and malign-game players' Stage 1 dictator-game givings, and these giving decisions have already been made in Stage 3 when the subject makes their choice of who to receive givings from. In fact, in Stage 3 they do not interact with their partners again. What subjects should do is to update their beliefs based on what happens in Stage 2 and choose the dictator-game givings of the player whom they believe to be more prosocial.

The comparison between Treatment 3 and Treatment 4 provides a further, direct, test of the preference for the good-experience partner explanation of the results. If behavior in Treatment 2 and Treatment 3 is driven by this effect, then we should see a similar size of correspondence bias in Treatment 4 compared to Treatment 3. In Treatment 4, subjects also play with one benign-game partner and one malign-game partner, which is no different from Treatment 3 in terms of experiences. If subjects' choices are driven by a preference to interact again with a partner they shared a positive experience with, then they should still be willing to pay more for benign-game partner's Stage 1 dictator-game transfers in Treatment 4. But what we observe is that subjects' WTP for benign-game partner's dictator-game transfers is no longer significantly different from 0 after receiving counterfactual information in Treatment 4, while the difference in the benign premium between Treatment 3 and Treatment 4 is positive and significant. This supports the interpretation that subjects are willing to pay a benign premium because they have a biased belief about the prosociality of the benign-game player, and not because they have a preference for the benign-game player after sharing a good experience.

Applications of Correspondence Bias

Correspondence bias has a wide range of applications in managerial decision-making, especially when it comes to hiring decisions. Employers constantly need to assess (potential) employees' ability based on their past achievements, which are the joint product of their abilities and effort, on the one hand, and the difficulty of the tasks they have been given and environments they have been placed in, on the other. Correspondence bias implies that employers tend to underestimate the influence of the task and environmental factors. Recent literature shows that graduating (and expecting to graduate) in a recession can have a long-lasting effect on people's earnings, employment, and health outcomes

(Kahn, 2010; Oreopoulos et al., 2012). The negative effect on earnings lasts ten years on average for unlucky graduates, and the disadvantaged ones among them may suffer a permanent loss. Correspondence bias could contribute to the strength and persistence of the effect, in addition to the usual accounts based on human capital accumulation. If employers are subject to correspondence bias, then they will be less likely to hire recession graduates, who, if they do find work, tend to work for smaller, less prestigious and lower-paying companies (Oreopoulos et al., 2012). Correspondence-biased employers will over-attribute initial labor market outcomes to employee's abilities while underestimating the impact of the labor market condition upon entry.

Interviews, one of the most commonly used tools in hiring, could also suffer from correspondence bias (Schmidt and Hunter, 1998). Even though many cognitive and non-cognitive tests have been developed to assess the productivity of job candidates, many employers still rely on one-shot, unstructured, face-to-face interviews to make their final hiring decisions. As many of those interviews are unstructured and only last less than an hour, the situations the job candidates face could vary dramatically across interviews. The interviewer could be in a good mood, or in a bad mood; the interviewer could ask some difficult questions, or some easy questions. Interviewees' performances in the interview could be the joint product of their innate ability and the circumstances they face. A correspondence-biased interviewer overly attributes the interviewee's performance to ability, and under-attributes it to environmental factors, resulting in error-prone evaluations that are given too much weight. This prediction is supported by the findings that only 14% of differences in employee productivity can be explained by interviews (Schmidt and Hunter, 1998), and managers who overrule objective job test results with subjective judgments based on interviews and other sources end up with worse hires on average (Hoffman et al., 2018).

Correspondence bias may also come into play when universities make student admission decisions. When evaluating students, admissions officers should, normatively, take account not only of the student's grade point average, but also the average grades of students in the university they are coming from – i.e., the stringency of grading. However, correspondence bias predicts that admissions officers will be excessively influenced by the former, and insufficiently by the latter, a prediction supported by research on admissions by Moore et al. (2010). Students who are from a college with a higher average GPA are more likely to be admitted by a graduate school compared to students with similar within-school rankings but who are from a comparable college of a lower average GPA. Swift

et al. (2013) further shows that informing admission experts of the distribution of GPAs in different colleges does not eliminate the effect of grade inflation, which indicates that the problem is not simply driven by experts' unawareness of the differences in grading leniency among colleges. More broadly, students' achievements in college are affected by a range of external factors beyond their control. Correspondence-biased admission officers may underestimate the impact of those factors, leading them to admit more privileged students who have achieved more in the past as a result of the opportunities they have been given, as opposed to their innate drive or intelligence. Although certainly not the only cause, correspondence bias may contribute to dramatically higher rates of admission to elite universities of students whose family incomes are in the top of the income distribution (Chetty et al., 2017).

In the domain of corporate governance, standard economic theory suggests that when evaluating the quality of a CEO the board should ignore firm-performance relevant factors that are out of the control of the CEO. However, empirical evidence shows that luck plays an important role in a CEO's career life cycle. Oil CEOs are rewarded for oil price increases that they have no role in creating (Davis and Hausman, 2020), import-affected sector's CEOs' pay is responsive to exchange rate changes (Bertrand and Mullainathan, 2001), and industry-level market shocks can affect CEOs' compensations (Bertrand and Mullainathan, 2001; Garvey and Milbourn, 2006). A lucky CEO not only earns more in her current firm but also enjoys better outside options (Amore and Schwenen, 2020). Meanwhile, an unlucky CEO is more likely to lose her job than her lucky counterparts for the same level of performance relative to the sector of her business (Jenter and Kanaan, 2015). Consistent with correspondence bias, CEOs are overly rewarded for good performance when the whole industry is doing well, and, even though boards of directors do recognize that performance in a downturn is more informative of a CEOs' abilities, they do not reward CEOs sufficiently for performing well in a downturn (Jenter and Kanaan, 2015), much as participants in our experiment were not willing to pay more for the transfer of a malign-game player who chose to cooperate than that of a benign-game player who chose to cooperate.

Luck can also affect the electoral prospects of politicians. Similar to oil CEOs, incumbent governors in oil-producing states are more likely to win re-election when oil prices increase, even though the international oil price is out of the governors' control (Wolfers, 2002). One potential explanation is that voters misattribute good or bad economic conditions that are partially driven by exogenous

factors to incumbents' abilities.

7 Conclusion

This paper investigates people's tendency to underestimate the influence of immediate incentives when making sense of others' behavior. The key intuition is that failing to fully appreciate the impact of incentives on actions leads individuals to over-attribute others' behaviors to their enduring characteristics.

We test the predictions of the model in an experiment with 817 subjects. We first ask subjects to decide how much to transfer as the dictator in a dictator game. Then, we let them play the benign game and the malign game, and inform them of the actions of a benign-game player and a malign-game player. Lastly, we ask them to choose, as a receiver in the dictator game, between the benign-game player and the malign-game player's first-stage transfers. In the baseline treatment, subjects experience one game but only observe a player's action in the other game, a situation that is probably most similar to those prevailing in real world situations – in which we interpret, and respond to, the behaviors of different people with only limited experiences of the situations they are in. We find strong evidence of correspondence bias. Subjects are willing to pay 12 cents out of a dollar to receive the benign-game player's dictator-game givings, which is significantly larger than what the Bayesian model predicts, 0. Allowing subjects to experience both games instead of playing one and observing the other one reduces correspondence bias, but the benign premium is still significantly above 0. However, if we inform subjects of how their benign-game player behaves in the malign game and vice versa, correspondence bias disappears.

Results from treatments 3 and 4 suggest that correspondence bias is caused by ignorance of the effect of incentives on actions. In Treatment 1, we directly test the predictions of our model: correspondence bias is driven by both overinference about prosociality from cooperation in the benign game and underinference about prosociality from cooperation in the malign game. We find that when choosing between a benign-game player and a random stranger subjects are on average willing to pay more for the benign-game player; when choosing between a malign-game player and a stranger, they are on average willing to pay more for the stranger.

Our findings shed light on why correspondence bias is widely observed in real life, as well as on potential ways to reduce or eliminate it. First, in reality, we often only experience one environment

and observe other environments, which makes it hard for us to understand how alternative environments affect other people's behaviors. This may help to explain research supporting the 'contact hypothesis' – showing that social cohesion is enhanced by encouraging social interactions between different groups (Rao, 2019; Lowe, 2020). Second, counterfactual information about how the people we encounter behave in other environments is rarely available; the broader the range of situations in which we observe another person, the current research suggests, the more we are likely to appreciate how contingent the individual's behavior is on the situation they are in.

REFERENCES

Ambuehl, Sandro and Shengwu Li, "Belief updating and the demand for information," *Games and Economic Behavior*, 2018, *109*, 21–39.

Amore, Mario Daniele and Sebastian Schwenen, "The value of luck in the labor market for CEOs," *CEPR Discussion Paper No. DP14839*, 2020.

Barreda-Tarrazona, Iván, Ainhoa Jaramillo-Gutiérrez, Marina Pavan, and Gerardo Sabater-Grande, "Individual characteristics vs. experience: an experimental study on cooperation in prisoner's dilemma," *Frontiers in Psychology*, 2017, *8*, 596.

Bednar, Jenna, Yan Chen, Tracy Xiao Liu, and Scott Page, "Behavioral spillovers and cognitive load in multiple games: An experimental study," *Games and Economic Behavior*, 2012, *74* (1), 12–31.

Benjamin, Daniel J, "Errors in probabilistic reasoning and judgment biases," in "Handbook of Behavioral Economics: Applications and Foundations 1," Vol. 2, Elsevier, 2019, pp. 69–186.

Bertrand, Marianne and Sendhil Mullainathan, "Are CEOs rewarded for luck? The ones without principals are," *The Quarterly Journal of Economics*, 2001, *116* (3), 901–932.

Bó, Ernesto Dal, Pedro Dal Bó, and Erik Eyster, "The demand for bad policy when voters underappreciate equilibrium effects," *The Review of Economic Studies*, 2018, *85* (2), 964–998.

Cason, Timothy N, Sau-Him Paul Lau, and Vai-Lam Mui, "Prior interaction, identity, and cooperation in the Inter-group Prisoner's Dilemma," *Journal of Economic Behavior and Organization*, 2019.

Charness, Gary, Luca Rigotti, and Aldo Rustichini, "Social surplus determines cooperation rates in the one-shot Prisoner's Dilemma," *Games and Economic Behavior*, 2016, *100*, 113–124.

Chetty, Raj, John N Friedman, Emmanuel Saez, Nicholas Turner, and Danny Yagan, "Mobility report cards: The role of colleges in intergenerational mobility," *NBER Working Paper*, 2017.

Choi, Incheol and Richard E Nisbett, "Situational salience and cultural differences in the correspondence bias and actor-observer bias," *Personality and Social Psychology Bulletin*, 1998, *24* (9), 949–960.

Cooper, Russell, Douglas V DeJong, Robert Forsythe, and Thomas W Ross, "Cooperation without reputation: Experimental evidence from prisoner's dilemma games," *Games and Economic Behavior*, 1996, *12* (2), 187–218.

Davis, Lucas W and Catherine Hausman, "Are energy executives rewarded for luck?," *The Energy Journal*, 2020, *41* (6).

Edwards, Ward, "Conservatism in human information processing," in Kleinmuntz B, ed., *Formal Reresentation of Human Judgement*, New York: Wiley, 1968, p. 17–52.

Esponda, Ignacio and Emanuel Vespa, "Hypothetical thinking and information extraction in the laboratory," *American Economic Journal: Microeconomics*, 2014, *6* (4), 180–202.

_ and _, "Contingent preferences and the sure-thing principle: Revisiting classic anomalies in the laboratory," *Working Paper*, 2019.

Eyster, Erik and Matthew Rabin, "Cursed equilibrium," *Econometrica*, 2005, *73* (5), 1623–1672.

Garvey, Gerald T and Todd T Milbourn, "Asymmetric benchmarking in compensation: Executives are rewarded for good luck but not penalized for bad," *Journal of Financial Economics*, 2006, *82* (1), 197–225.

Gawronski, Bertram, "Theory-based bias correction in dispositional inference: The fundamental attribution error is dead, long live the correspondence bias," *European Review of Social Psychology*, 2004, *15* (1), 183–217.

Gilbert, Daniel T, "Thinking lightly about others: Automatic components of the social inference process," *Unintended Thought*, 1989, *26*, 481.

_ **and Patrick S Malone**, "The correspondence bias.," *Psychological Bulletin*, 1995, *117* (1), 21.

Graeber, Thomas, "Inattentive inference," *Working Paper*, 2020.

Haggag, Kareem, Devin G Pope, Kinsey B Bryant-Lees, and Maarten W Bos, "Attribution bias in consumer choice," *The Review of Economic Studies*, 2019, *86* (5), 2136–2183.

_ **, Richard W Patterson, Nolan G Pope, and Aaron Feudo**, "Attribution bias in major decisions: Evidence from the United States Military Academy," *IZA Discussion Paper*, 2019.

Hoffman, Mitchell, Lisa B Kahn, and Danielle Li, "Discretion in hiring," *The Quarterly Journal of Economics*, 2018, *133* (2), 765–800.

Jenter, Dirk and Fadi Kanaan, "CEO turnover and relative performance evaluation," *The Journal of Finance*, 2015, *70* (5), 2155–2184.

Jones, Edward E and Victor A Harris, "The attribution of attitudes," *Journal of Experimental Social Psychology*, 1967, *3* (1), 1–24.

Kahn, Lisa B, "The long-term labor market consequences of graduating from college in a bad economy," *Labour Economics*, 2010, *17* (2), 303–316.

Kahneman, Daniel and Amos Tversky, "Subjective probability: A judgment of representativeness," *Cognitive Psychology*, 1972, *3* (3), 430–454.

_ **and** _ , "On the psychology of prediction.," *Psychological Review*, 1973, *80* (4), 237.

Lowe, Matt, "Types of contact: A field experiment on collaborative and adversarial caste integration," *CESifo Working Paper*, 2020.

Martínez-Marquina, Alejandro, Muriel Niederle, and Emanuel Vespa, "Failures in contingent reasoning: The role of uncertainty," *American Economic Review*, 2019, *109* (10), 3437–74.

Mengel, Friederike, "Risk and temptation: A meta-study on Prisoner's Dilemma games," *The Economic Journal*, 2018, *128* (616), 3182–3209.

Möbius, Markus M, Muriel Niederle, Paul Niehaus, and Tanya S Rosenblat, "Managing self-confidence," *NBER Working Paper*, 2014.

Moore, Don A, Samuel A Swift, Zachariah S Sharek, and Francesca Gino, "Correspondence bias in performance evaluation: Why grade inflation works," *Personality and Social Psychology Bulletin*, 2010, *36* (6), 843–852.

Morewedge, Carey K, Haewon Yoon, Irene Scopelliti, Carl W Symborski, James H Korris, and Karim S Kassam, "Debiasing decisions: Improved decision making with a single training intervention," *Policy Insights from the Behavioral and Brain Sciences*, 2015, *2* (1), 129–140.

Ngangoué, M Kathleen and Georg Weizsäcker, "Learning from unrealized versus realized prices," *American Economic Journal: Microeconomics*, 2021, *13* (2), 174–201.

Oreopoulos, Philip, Till Von Wachter, and Andrew Heisz, "The short-and long-term career effects of graduating in a recession," *American Economic Journal: Applied Economics*, 2012, *4* (1), 1–29.

Peysakhovich, Alexander and David G Rand, "Habits of virtue: Creating norms of cooperation and defection in the laboratory," *Management Science*, 2015, *62* (3), 631–647.

Phillips, Lawrence D and Ward Edwards, "Conservatism in a simple probability inference task," *Journal of Experimental Psychology*, 1966, *72* (3), 346.

Rao, Gautam, "Familiarity does not breed contempt: Generosity, discrimination, and diversity in Delhi schools," *American Economic Review*, 2019, *109* (3), 774–809.

Ross, Lee, "The intuitive psychologist and his shortcomings: Distortions in the attribution process," *Advances in Experimental Social Psychology*, 1977, *10*, 173–220.

Schmidt, Frank L and John E Hunter, "The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings.," *Psychological bulletin*, 1998, *124* (2), 262.

Sherman, Steven J, "On the self-erasing nature of errors of prediction.," *Journal of personality and Social Psychology*, 1980, *39* (2), 211.

Simonsohn, Uri, Niklas Karlsson, George Loewenstein, and Dan Ariely, "The tree of experience in the forest of information: Overweighing experienced relative to observed information," *Games and Economic Behavior*, 2008, *62* (1), 263–286.

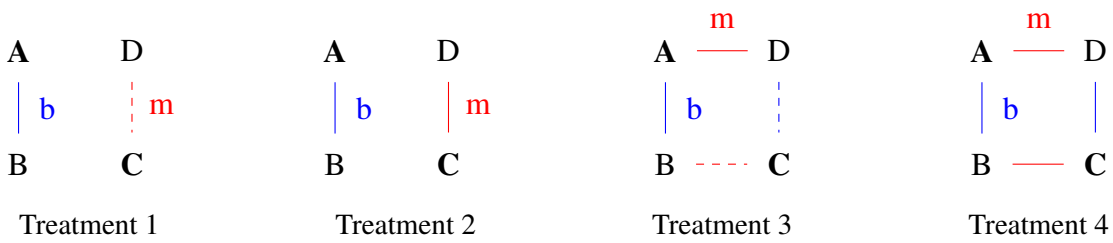
Swift, Samuel A, Don A Moore, Zachariah S Sharek, and Francesca Gino, "Inflated applicants: Attribution errors in performance evaluation by professionals," *PLoS One*, 2013, *8* (7), e69258.

Walker, Drew, Kevin A Smith, and Edward Vul, "The 'Fundamental Attribution Error' is rational in an uncertain world," *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, 2015.

Wolfers, Justin, "Are voters rational? Evidence from gubernatorial elections," *NBER Working Paper*, 2002.

Figures

Figure 1: Overview of Four Treatments

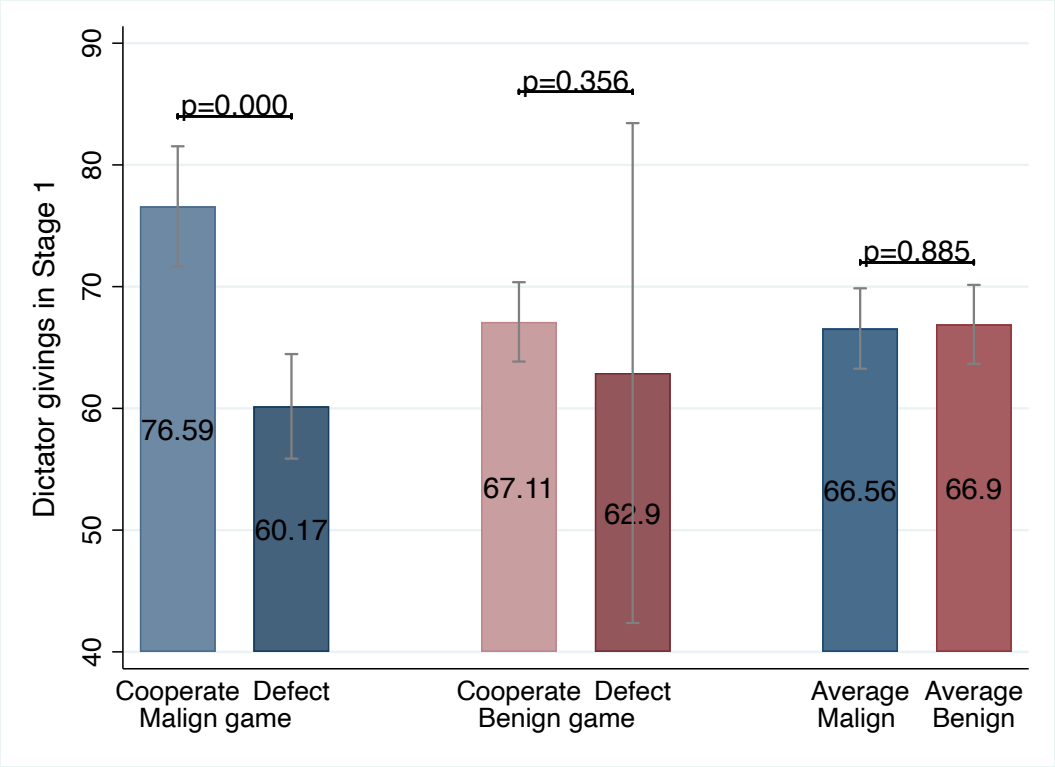


b: benign game m: malign game

: observed by A :not observed by A

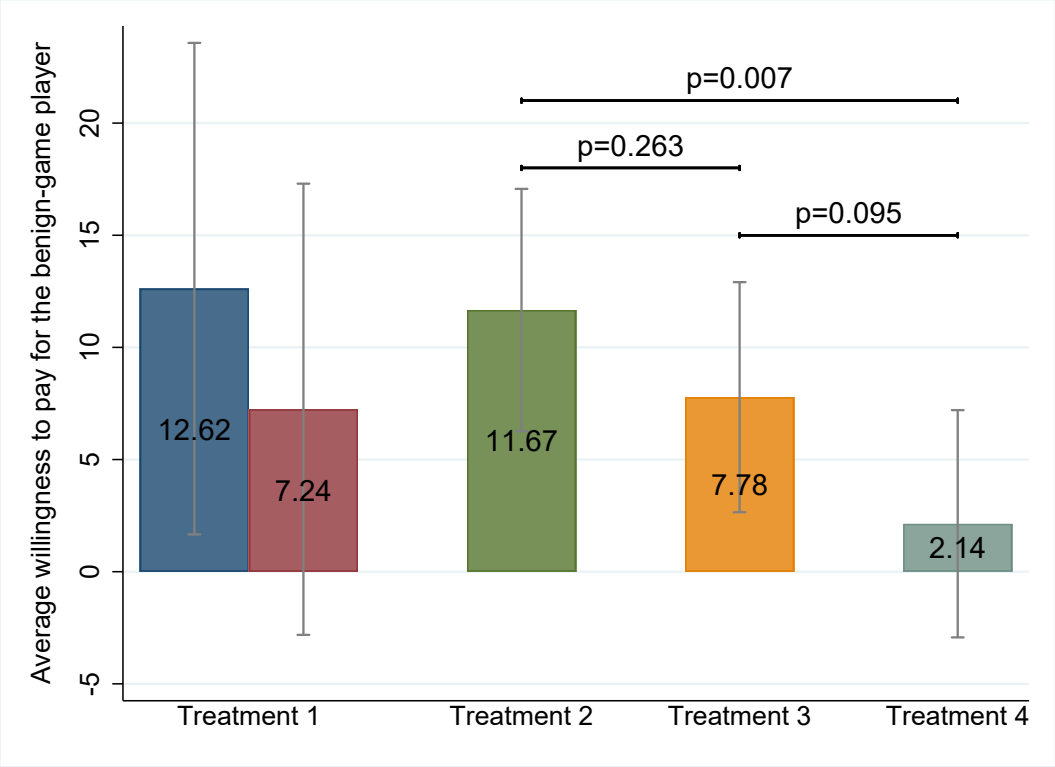
Notes: The figure displays the four treatments from subject A's perspective. The solid line denotes that A is able to observe (the outcome of) a game, and the dashed line denotes that A is not able to observe a game. But of course, A is not the only active player in the game. The games faced by B, C, and D are symmetric in treatments 2, 3, and 4. For example, player D plays the benign game with C and the malign game with A in Treatment 3. She cannot observe the game played between A and B or the game played between C and B in Treatment 3. The game is not symmetric in Treatment 1. In that treatment, A and B only play the benign game, and C and D only play the malign game.

Figure 2: Dictator Givings as a Function of Action in the Malign Game



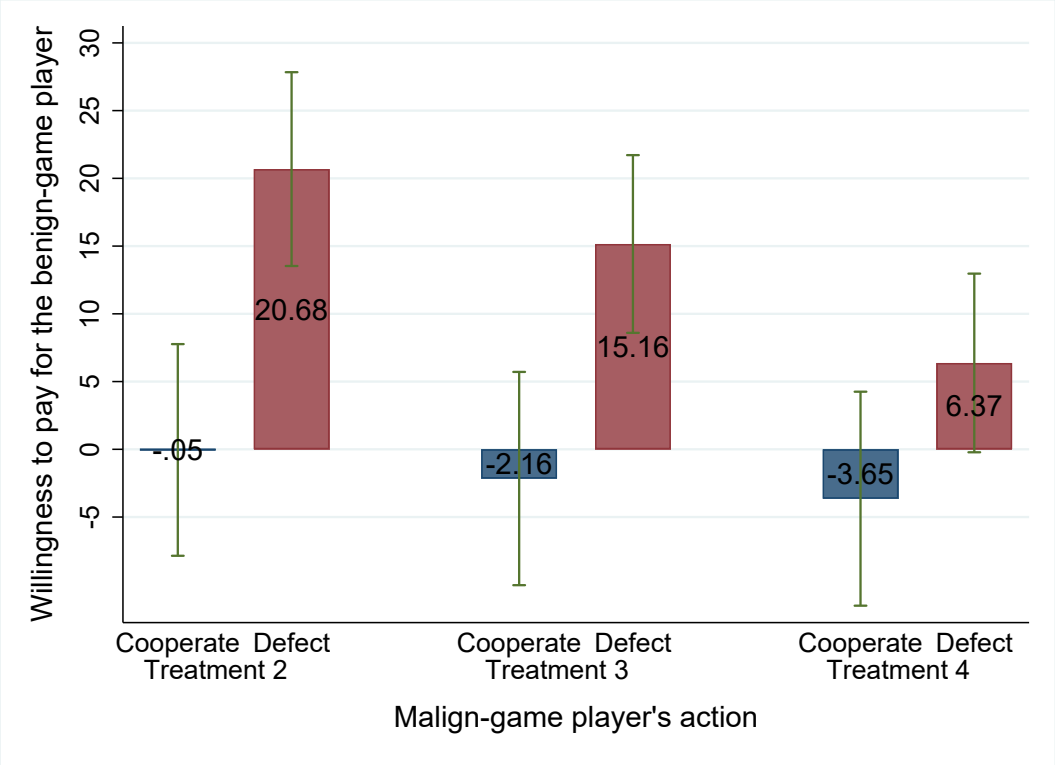
Notes: The figure plots the average dictator givings in the dictator game in Stage 1 depending on their own actions in the malign game. Recall that subjects were asked to divide 200 cents between themselves and a random receiver in the first stage. The bars show means of dictator givings, and the vertical lines report 95% confidence intervals.

Figure 3: Benign Premiums across Treatments



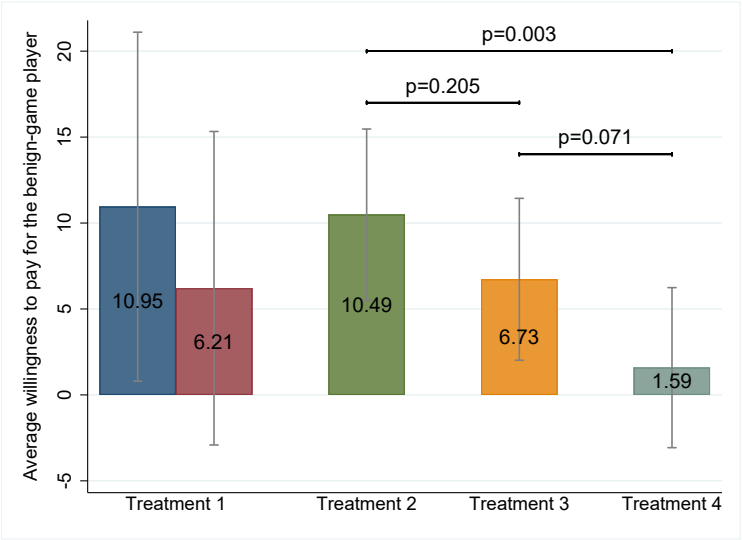
Notes: The figure plots the benign premiums across treatments. The bars show means of WTP for a benign-game player in different treatments. The vertical lines report 95% confidence intervals. The left bar in Treatment 1 represents the average WTP for a benign-game player when choosing between her and a stranger, and the right bar in Treatment 1 represents the average WTP for a stranger when choosing between him and a malign-game player.

Figure 4: Average WTP for Benign-game Player as a Function of the Malign-game Player's Action

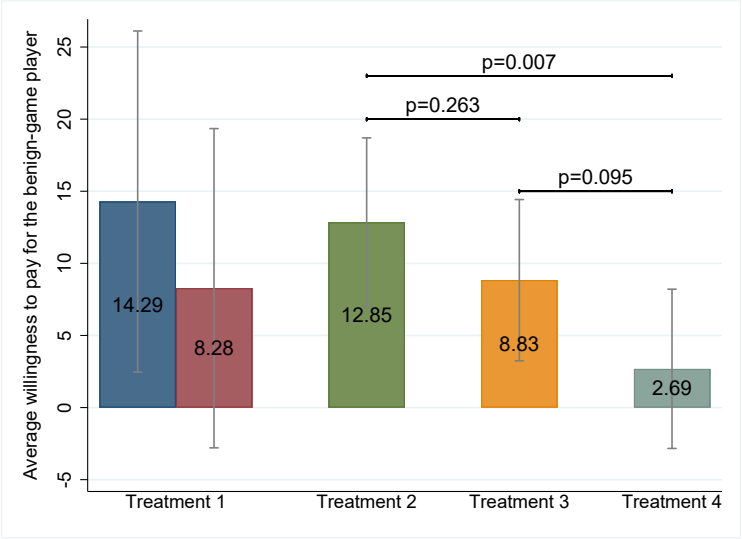


Notes: The figure plots the average WTP for a benign-game player in treatments 2 to 4, depending on their malign-game player's action. The bars show means of WTP and the vertical lines report 95% confidence intervals.

Figure 5: Robustness Check - Benign Premium across Treatments



Panel (a)



Panel (b)

Notes: The figure plots the average WTPs towards the benign-game player over the four treatments using alternative coding methods. In our main analysis, we code the WTP as the median value of the interval at which subjects switched from one option to the other. In Panel (a), WTP is defined instead as the lower bound of that interval. In Panel (b), WTP is defined as the upper bound of that interval. In both panels, the left bar in Treatment 1 represents the average WTP for a benign-game player when choosing between her and a stranger, and the right bar in Treatment 1 represents the average WTP for a stranger when choosing between him and a malign-game player.

Tables

Table 1: The Benign and Malign Games

Harmony Game

	C	D
C	40,40	10,30
D	30,10	0,0

Prisoner's Dilemma

	C	D
C	40,40	20,120
D	120,20	30,30

Notes: The harmony game is the benign game, and the prisoner's dilemma is the malign game.

Table 2: The Multiple Price List

	Option 1	Option 2
Choice 1	Amount transferred to me by B	Amount transferred to me by D+¢10
Choice 2	Amount transferred to me by B	Amount transferred to me by D+¢20
Choice 3	Amount transferred to me by B	Amount transferred to me by D+¢30
Choice 4	Amount transferred to me by B	Amount transferred to me by D+¢40
Choice 5	Amount transferred to me by B	Amount transferred to me by D+¢50
Choice 6	Amount transferred to me by B	Amount transferred to me by D+¢60
Choice 7	Amount transferred to me by B	Amount transferred to me by D+¢70
Choice 8	Amount transferred to me by B	Amount transferred to me by D+¢80
Choice 9	Amount transferred to me by B	Amount transferred to me by D+¢90
Choice 10	Amount transferred to me by B	Amount transferred to me by D+¢100

Notes: The table shows the multiple price list shown to subject A if she chose B over D in the first choice.

Table 3: Summary Statistics

Variable	All Sample	Treatment One	Two	Three	Four
Dictator giving	67.42	71.12	69.13	65.72	65.26
	(41.60)	(39.73)	(40.77)	(42.47)	(42.64)
Cooperation rate in the benign game	0.952	0.921	0.969	0.942	0.960
	(0.215)	(0.272)	(0.175)	(0.235)	(0.196)
Cooperation rate in the malign game	0.389	0.397	0.420	0.359	0.401
	(0.488)	(0.493)	(0.496)	(0.481)	(0.491)
Survey completion rate	0.903	0.901	0.902	0.897	0.912
	(0.296)	(0.300)	(0.297)	(0.305)	(0.284)
Observations	817	121	246	223	227
Follow-up survey					
Income	3.861	3.815	3.914	3.864	3.826
	(1.599)	(1.486)	(1.648)	(1.549)	(1.657)
Female	0.574	0.556	0.584	0.623	0.527
	(0.495)	(0.499)	(0.494)	(0.486)	(0.501)
Age	38.04	37.44	38.43	37.13	38.81
	(10.94)	(10.09)	(11.41)	(9.882)	(11.79)
Employment	0.819	0.824	0.819	0.829	0.807
	(0.385)	(0.383)	(0.386)	(0.377)	(0.396)
Observations	735	108	221	199	207

Notes: The table reports the mean for each variable in the whole sample and across treatments, with standard deviations in parentheses. We collect subjects' demographic information in a voluntary follow-up survey. 735 out of 817 subjects completed the survey. Income is a categorical variable, with categories 1="Less than $25,000", 2="$25,000 to $34,999", 3="$35,000 to $49,999", 4="$50,000 to $74,999", 5="$75,000 to $99,999", 6="$100,000 or more." Employment is defined as the percentage of people who are currently self-employed or employed.

Table 4: Cooperation and Dictator Givings by Games

	Benign game	Malign game
Cooperation rate	0.952	0.389
	(0.215)	(0.488)
Dictator givings	66.90	66.56
	(41.76)	(42.11)
Obs	640	627

Notes: The table reports the average dictator givings (in cents) and cooperation rates in the two games, with standard deviations in parentheses.

Table 5: Benign Premiums across Treatments

Treatment		Obs	Benign Premium	P-value $H_0 : BP = 0$	P-value $H_0 : BP_{Tx} = BP_{T2}$
Treatment 1	benignP VS stranger	63	12.62	0.025	
	stranger VS malignP	58	7.24	0.155	
Treatment 2		246	11.67	0.000	
Treatment 3		223	7.78	0.003	0.263
Treatment 4		227	2.14	0.407	0.007

Notes: The first row in Treatment 1 represents the average WTP for a benign-game player when choosing between the benign-game player and a random stranger, and the second row in Treatment 1 represents the average WTP for a stranger when choosing between the stranger and a malign-game player. BP stands for benign premium. Column (3) reports the p-value of t-tests, and column (4) reports the p-value of rank-sum tests.

Table 6: Benign Premiums and Fractions across Treatments and Malign-game Player's Actions

			Malign-game player				Rank-sum test	
			Cooperate		Defect		p-value	
	Fraction (1)	BP (2)	Fraction (3)	BP (4)	Fraction (5)	BP (6)	(3)vs(5)	(4)vs(6)
Panel A								
Treatment 2	0.62	11.66	0.52	-0.05	0.69	20.68	0.008	0.000
	(0.49)	(43.00)	(0.50)	(40.76)	(0.46)	(42.65)		
Obs	246	246	107	107	139	139		
BenignG only	0.74	21.93	0.74	13.89	0.74	27.88	0.990	0.042
	(0.44)	(40.78)	(0.44)	(35.91)	(0.44)	(43.32)		
Obs	127	127	54	54	73	73		
MalignG only	0.49	0.71	0.30	-14.25	0.64	12.73	0.000	0.000
	(0.50)	(42.78)	(0.46)	(40.80)	(0.48)	(40.75)		
Obs	119	119	53	53	66	66		
Panel B								
Treatment 3	0.61	7.78	0.52	-2.16	0.67	15.16	0.019	0.002
	(0.49)	(38.86)	(0.50)	(38.64)	(0.47)	(37.50)		
Obs	223	223	95	95	128	128		
Panel C								
Treatment 4	0.56	2.14	0.47	-3.65	0.62	6.37	0.025	0.013
	(0.50)	(38.73)	(0.50)	(38.98)	(0.49)	(38.14)		
Obs	227	227	96	96	131	131		

Notes: The table shows the fractions of subjects who chose the benign-game player over the malign-game player (*Benign Fraction*) in choice 1 with no bonuses and the benign premiums in treatments 2, 3, and 4. BP stands for the benign premium and Fraction stands for the benign fraction. Standard deviations are in parentheses. Columns 1 and 2 report the benign fractions and benign premiums in the three treatments respectively. Columns 3,4 and Columns 5,6 report the same statistics when the malign-game player chose to cooperate and defect respectively. Column 7 presents the p-value of a rank-sum test that the mean levels are the same for columns 3 and 5; column 8 presents the same test for columns 4 and 6. In Panel A, *BenignG only* denotes subjects who played the benign game and observed the malign game; *malignG only* denotes subjects who played the malign game and observed the benign game.

Table 7: Heterogeneous Analysis

Dependent Variable	Benign Premiums		
	All	Survey	Survey
	(1)	(2)	(3)
Treatment 3	-2.959	-4.730	-4.132
	(3.764)	(4.016)	(3.925)
Treatment 4	-8.430**	-8.004**	-7.888**
	(3.812)	(4.034)	(3.952)
Education			1.069
			(1.911)
Stage 1 stay time			0.032
			(0.045)
Stage 2 stay time			0.010
			(0.035)
Stage 2 results stay time			0.008
			(0.012)
Stage 3 stay time			-0.006
			(0.007)
Observations	696	627	627

Notes: The table reports results from an interval regression to address the concern that multiple price list only elicits intervals of WTP. Observations are subjects in treatments 2, 3, and 4. The omitted group is Treatment 2. Column 1 includes all subjects in treatments 2, 3, and 4. Columns 2 and 3 include subjects who completed the follow-up survey. All regressions include the date of participation fixed effects. In column 3, we also include subjects' gender, income, risk preference, malign-game player's action. *** $p<0.01$, ** $p<0.05$, * $p<0.1$. Standard errors in parentheses.