2015

# The Strategy of Explanations

Sean McCoy

# The Strategy of Explanations

Sean McCoy[*]

24 April 2015

**Abstract**

To better understand the strategic aspects of human decision-making, we conduct a power-to-take game in which takers are required to send messages explaining their actions to receivers. We vary the types and timing of information takers have available to them to investigate the effects of empathy and deceptive framing. We determine that takers who have more information tend to act and explain themselves more strategically. We also find that receivers prefer to interact with female takers even though all interactions are anonymous. These findings suggest there is value in explanations that is often overlooked in modern analyses of human decision-making.

Keywords: power-to-take game, altruism, empathy, stories, pro-social behavior

Stories have been an integral part of human discourse since there has been interest in recording people's thoughts and actions. They can be fabricated or truthful, purposeful or meaningless, and they can convey almost every human emotion. Consequently, we use them several times a day for a variety of reasons, and it is for these reasons that they have strategic value in decision-making. And,

[*] With Benjamin Ho, Department of Economics, Vassar College, Poughkeepsie, NY.
Email at: beho@vassar.edu and semccoy@vassar.edu respectively.
We give special thanks to Alexander Egan, Jonathan Maesato, Richard Winkeller, Paul Ruud, and Shereen Chaudhry for their help in developing this project, and we gratefully acknowledge the financial support of the Elinor Nims Vassar Research Fund and Vassar's Academic Enrichment Fund.

since we control the ways in which we express our stories, we can often adapt them to serve our interests.

For example, consider a job interview and the common question “Where do you see yourself in five years?” There are myriad ways to respond—and all are stories to tell—but some are more informative and influential than others. The applicant may employ one of several strategies to present himself in the best light because he wants to impress the interviewer, demonstrate interest in the firm, and emphasize certain aspects of his character. So perhaps the he says that he wants to be sitting in his interviewer’s shoes in five years. This is a vague response that does not satisfy any of the applicants’ goals and may even come across as an affront to the interviewer; this is not a good strategic option because the costs the story incurs outweigh its benefits. On the other hand, if the applicant lays out his goals along a realistic career path that is complemented by employment at the firm, he will have strategically used a story to his advantage. Although this example is trivial, its implications are quite broad. The story the applicant tells is costly in the sense that it has great influence on the first impression he makes with the interviewer, and thus it is a tool he can use to influence his wellbeing. Accordingly, he should treat it, and all other stories he tells, as such. Similarly, any story that is inappropriately valued may be used inefficiently because its effects are absent from the utility considerations of the person telling the story. In short, we must consider the strategic value of stories if we are to more completely understand the ways in which we interact and the optimal strategies for doing so.

Though the economic literature on stories is relatively sparse, others have studied similar forms of communication at length by observing interactions in the power-to-take game (Bosman and van Winden, 2002). Though we explain our particular version of the game in detail later on, we maintain the basic structure of the game by partnering one active player (the “taker”) with one passive player (the “receiver”) and bestowing the taker with the ability to take the receiver’s resources. Until recently, the receiver has been the main focus of much of the extant research on the power-to-take game and interpersonal interactions (Bosman et al., 1999, 2000, 2005; Ben-Shakhar et al., 2007; Grosskopf and López-Vargas, 2014), whereas less has been done on takers. And, while people's behavior and emotions can be attributed to mechanisms such as profit maximization, empathy (Andreoni and Rao, 2011), and guilt aversion (Charness and Dufwenberg, 2006), it is likely that the means of expressing those actions and emotions—that is, the stories themselves that confer them—play a considerable role in their reception. For example, Andreoni and Rao (2011) examine the role communication plays in influencing altruism by investigating how it affects behavior in a dictator game. They find that more altruistic behavior occurs when receivers ask takers to give them resources, that dictators who can explain their actions however they want are particularly selfish, and that empathy promotes altruism. We seek to expand directly on these results by varying the types and timing of information takers have available to them at different stages in their interactions. By treating empathy as a time-sensitive emotion, we stand to learn more about the strategy of human decision-making by honing in on the factors that contribute to it.

Grosskopf and López-Vargas (2014) explore receivers' desires to express their emotions and the value of ex-post verbal communication. They conduct a power-to-take game in which takers can, as usual, seize the assets of receivers, but add the restriction that receivers cannot stop this from happening. Additionally, some receivers are given the option to send either free or costly messages to their partners through neutral third parties while others cannot send messages at all. The authors find that receivers are willing to pay a significant amount of their earnings to express their emotions, especially when takers have been particularly generous or stingy with them. Along with this comes the finding that receivers whose moods are more strongly affected by experimental events tend to value their expressions more and are therefore more willing to pay to express them to their partners. They also find that takers behave drastically more altruistically when they are aware that the receivers they are playing with can send messages back to them, suggesting that takers who are expecting receivers to respond act less selfishly than those who know that takers cannot respond. This finding complements that of Ellingsen and Johannesson (2008), who investigate the role that anticipated verbal feedback has in determining altruistic behavior and find that expectations of correspondence lead to more altruistic behavior.

We also examine the emotions of receivers in our experiment, both to reinforce the existing literature and because understanding their feelings regarding takers should help us understand more about how future interactions between people may play out. Ben-Shakhar et al. (2007) investigate the effects of costly reciprocation and emotional ties between players in a standard power-to-take game. They find that players act less self-servingly and favor more equal distributions of resources when they know information about each other. They also note that expectations of behavior are reflected more strongly in changes in emotional state than in changes in the actual game being played; that is to say, people's responses are based heavily on preconceptions—perhaps both verbal and nonverbal cues of some sort—and to a lesser extent on how real events compare with those expectations. Additionally, Bosman and van Winden (2002) find that receivers' expectations of greed significantly affect their likelihood to destroy their own incomes so that takers cannot seize them, further supporting the notion that people evaluate both the monetary viewpoint from choosing a course of action and the resultant emotional viewpoint that an action may bring. Reuben and van Winden (2006) also explore the effects of time on the role of takers in repeated power-to-take games; they operate a game in which takers can choose to change their actions between rounds in response to receivers' feedback. The authors find that takers' feelings of shame, guilt, and perceptions of fairness shape their future interactions with receivers. It is in part for this reason that we repeat our experiment over three rounds, thereby giving takers the ability to reflect on their own actions, explanations, and those of their partners such that they have more information upon which to base future decisions.

In addition, there exists a body of research specific to the economics of expressing emotions in verbal, written, and other forms. Xiao and Houser (2005) find that negative emotions do have a quantifiable monetary value and that there is a significant demand for expressing emotions. Additionally, they note that costly, self-inflicted punishments—such as destroying one's own resources in a power-to-take game—may help people express their negative emotions. They also study (2007) dictator games in which takers seize sub-profit-maximizing amounts when receivers can send them ex-post written messages, suggesting a fear of judgment or retribution akin to that found in Grosskopf and Lopez-Vargas's recent work (2014). However, they also find that monetary threats prove more effective than stories when determining players' actions in repeated games. This suggests that either stories lose value over time or that monetary incentives increase over time; we try to separate out these two effects in our experiment.

Finally, in looking at the specific kinds of messages that people send, Ho (2012) finds that apologies are common early in relationships and prevalent in relationships between well-matched people, suggesting that apologies in our experiment should be used relatively sparingly since players are poorly matched. Thus if they are overused, it may be the case that takers are trying in vain to start relationships on the basis of apologies—and their apologies fall on deaf ears—because both players are not at the stage of a relationship where apologies are meaningful. Additionally, Charness and Dufwenberg (2006) find that people try to live up to others' expectations and that promises encourage trust and cooperation between players, so perhaps the payoffs from making promises, which are not always monetary, can nonetheless be significant in governing interpersonal relationships. To this end, Vanberg (2008) studies the effects of making of promises in interpersonal interactions and finds that, while people have an inherent preference for keeping their promises, they do not do so because promises guarantee specific future payoffs; instead, there is usually some other mechanism at work, but it tends to vary situationally. This suggests that, while telling a given story and behaving in a way that causes people to believe you are living up to it is important, there is more to people's actions than the stories they tell about them.

## I. Theory

To begin to answer these questions, we employ a power-to-take game, which is akin to the well-known dictator game (Kahneman et al., 1986; Forsythe et al., 1994; Andreoni and Miller, 2004). The payoffs in the power-to-take game are isomorphic to those of the dictator game, but the power-to-take game has one key difference, that the dictatorial "taker" party gets to allocate the assets of their partner, the "receiver," and not their own (Bosman and van Winden, 2002). This aspect of the game introduces a difference in how the payoffs are framed; a dollar taken is not the same as a dollar not given (List 2007). In the standard power-to-take game (Bosman and van Winden, 2002), there are two players and two stages. First, the taker, who initially earns income $Y_T$, selects some value $t \in [0,1]$, which is the proportion of the receiver's income to be transferred to the taker at the end of the game, which we call the "take rate." Next, the receiver, who initially has income $Y_R$, learns about the taker's chosen $t$ and selects some value $d \in [0,1]$, which is the proportion of their income $Y_R$ that will be destroyed prior to the taker seizing any of it. Thus final payoffs are $Y_T + t(1\text{-}d)Y_R$ for the taker and $(1\text{-}t)(1\text{-}d)Y_R$ for the receiver. If both parties are strictly profit maximizing, the taker will select $t = 1$ and the receiver will be indifferent between all levels of $d$, but if for some reason $t < 1$, $d$ will be 0 as any $d > 0$ avoidably harms the receiver. Thus there may be incentive for the taker to select $t << 1$ as they risk playing a profitless scenario if $t$ is high enough to cause the receiver to select $d = 0$. Accordingly, there is reason for takers to consider the feelings of their partners, for if they do not, their profits and wellbeing may suffer as a result (Galeotti, 2013).

However, this analysis is contingent upon the receiver's ability to select $d$; without it—their only source of power in the standard version of the game—nothing apart from not wanting to endure the receiver's response prevents the taker from maximizing profits by selecting the highest $t$ possible. And, in an anonymous scenario, there is ample reason for them to take this course of action. But we are interested in takers' behavior for reasons discussed below, and because of this, we impose a restriction found in Grosskopf and Lopez-Vargas (2014), that receivers cannot prevent takers from taking their resources. This liberates takers to some extent since they no longer have to worry about receivers selecting high values of $d$, and instead, they can focus on decisions that serve their own self-interests: Payoffs are now simplified to $Y_T + tY_R$ for the taker and $(1-t)Y_R$ for the receiver. With takers in this mindset, we introduce another restriction, that takers must explain their behavior to receivers when they select a value of $t$. These explanations, the "stories" in our experiment, allow us to study the reasoning behind takers' actions, which we can then use to better understand their decision-making processes. Ultimately, the taker's payoff becomes $Y_T + tY_R + s(T, R)$ where s is a function describing the costs and benefits the taker's explanation has for both players.

## II. Experimental design and mechanisms

We deliberately set up our experiment with an emphasis on the taker role because, as a result of their privileges described above, takers are allowed and perhaps even encouraged to behave in ways that are often seen as selfish. Consequently, it is likely that their explanations will be costly since they may find it uncomfortable to think about and relate their behavior. This is especially the case if they have behaved in a way that receivers could construe as selfish. So, at risk of implicating themselves, we expect takers to treat their explanations as strategic stories that may harm their wellbeing if misused. However, this also gives rise to an alternative for takers, one that we suspect is a major mechanism motivating human decision making: *deceptive framing*.

For our purposes, we define deceptive framing as an attempt to affect the reception of a story by altering the way in which it is told. Lying, for example, would fall under this category because it obfuscates the true emotions that are masked by the words in a given story; what is said is not meant, and the intention of this incongruity is to better the person telling the story. This strategy should theoretically be common for two reasons: First, we often avoid the truth as a means of coping or self-defense and second, we frequently act with the intent to deceive others. Because of these tendencies, it is feasible that takers would rather lie to get out of uncomfortable explanations than mentally relive and physically recount their behaviors truthfully, especially if they are worried about their actions being frowned upon. However, we hope to determine whether interactions are generally predicated by the intent to deceive for personal gain or whether doing so is a byproduct of some other force at work. In our experiment, we evaluate the taker's explanation, the taker's behavior, and the receiver's reception of the takers' actions to evaluate whether or not the taker has employed a strategy of deceptive framing.

Also, much like Andreoni and Rao, we consider *empathy*, as well as the timing of empathetic feelings, to be another important mechanism that drives decision-making. In their experiment, communication was allowed between the dictator and the receiver, and this heightened feelings of empathy—and altruism—as the dictator often found himself considering the scenario from the point of view of the receiver. In our case, this is subsumed by the function s(T, R) in a taker's payoff, and we expect empathy to be most evident in comparisons between takers' explanations and actions. It has been suggested in the empathy-altruism hypothesis (Batson et al., 1988) that the desire to feel empathy, and not the desire to benefit oneself, is the main reason that people partake in altruistic behavior. This belief, in conjunction with our proposition that takers treat stories strategically, suggests that feelings of empathy should be evident in generous takers' explanations but largely absent from those of stingy ones. Additionally, we provoke feelings of empathy in takers by informing them of the role receivers will play in the experiment. This serves both to heighten takers' awareness of their partners' conditions and to cause takers to see the costs of their own explanations from receivers' points of view. Accordingly, we expect takers in treatments that know more about receivers to behave more altruistically in an effort to better themselves by not disadvantaging receivers.

Additionally, since people make decisions that govern interactions at roughly the same time that they interact—i.e., their stories are generally not planned well in advance—this setup allows us to investigate people's preferences for time-dependent utility. For instance, it may be difficult to overcome the impulsive desire to explain oneself in the first way that comes to mind, and this difficulty may lead to suboptimal decisions being made out of convenience. Also, people's preferences for utility may differ depending on their behavior or the contents of their stories; it is not unlikely that those who behave altruistically have a longer-term view of the benefits their actions might have for others or that those who use their stories to achieve a given strategy think about interactions in a fundamentally different way than others.

Finally, we stress the social significance of an interest in the taker role again because of an interest in timing: People often think and act to serve their own interests before considering how their actions affect others. Set in the context of our experiment, people are more likely to be takers than receivers in everyday interactions. For example, if someone performs an action that merits a costly explanation, they become the taker in that scenario. Accordingly, all of the conversations that we initiate are in some way a manifestation of the taker role, and so long as we choose how we explain ourselves, we can strategically direct the course of conversation and therefore have some amount of control over our own payoffs. Thus studying takers in particular provides us with useful information that can help us understand how people think about interpersonal interactions.

### A. Taker experiment

Before completing the experiment, all participants were given written instructions, all of which are available in the appendix. They then answered standard demographics questions and completed an effortful task, which entailed answering four questions about a simple data table. We paid them $\pi \in \{\$0.20, \$0.40, \$0.80\}$ for this task to avoid endowment effects. Next, participants were anonymously paired in groups of two. The taker then had the option to take any amount $t \in \{\$0.00, \$0.01, \dots \$0.20\}$ from the receiver. The receiver could not prevent this loss; they were required to accept the division of earnings as divided by the taker. Takers then had to explain their behavior. Over the course of the experiment, takers earned each possible value of □ listed above exactly once, but, in an effort to encourage taker recalibration between rounds, they did not know which round of the experiment would yield which payoff.

Next, we elicited a willingness to pay (wtp) or willingness to accept (wta) from select takers by flipping a two-sided coin. If the coin landed heads, their explanation would have been sent to their partner by default, but they could have opted to pay any amount $wtp \in \{\$0.00, \$0.01, \ldots \$0.20\}$ to try to prevent it from being sent. We then generated a random number $r$ in the range of $wtp$, and if $wtp \geq r$, the message was not sent; otherwise, it was. In both cases, the taker forfeited the value $r$ for attempting to change the course of events. Similarly, if the coin landed tails, the message was not sent by default, and the taker could have selected a $wta \in \{\$0.00, \$0.01, \ldots \$0.20\}$ to try to guarantee that the message would be sent. A taker could not have both a positive wtp and wta; only one was allowed to exceed $0.00. We elicited both wtps and wtas so as not to restrict takers' choices regarding how their explanations were conveyed.

This concluded one round of the experiment, and two more rounds followed. Takers played against the same receivers in subsequent rounds, but they knew them by different aliases, so they had no reason to believe that they were playing against the same people over and over. Both players retained their roles in all three rounds. No information was divulged between rounds, but takers were reminded of their and their partner's privileges each round. Takers could therefore earn up to $\pi + t - w$ per round, where $w$ represents wtp or wta, whichever, if either, the taker preferred. In total, takers could earn up to $\Pi = \sum_{i=1}^{3} (\pi_i + t_i - w_i) \quad \in \{\$0.00 \ldots \$2.80\}$.

The messages we collected from each participant were sorted using MTurk. First, we asked three respondents to describe each message with a single adjective. Next, we formed five main categories—apology, fairness, guilt, honesty, and selfishness—by grouping together the most popular adjectives and their derivatives (e.g. the “apology” category would encompass terms like “apology,” “apologetic,” “sorry,” etc.). Then, we asked five different respondents to sort each message into one of the five aforementioned categories or a “none of the above” category if they felt none were satisfactory. Messages were categorized by majority vote; messages with 3 apology, 1 guilt, and 1 honesty votes were “apology” messages, whereas messages with 2 fairness, 1 apology, and 2 selfishness votes went uncategorized because they were not decidedly one type of message or another. Nearly 95% of all messages were sorted into one of these categories by majority vote.

Note that we restrict messages to one category because categorization becomes extraordinarily tedious otherwise. While it is clear that certain messages contain elements that could group them in two or more categories, we find that partial categorization adds nothing significant to our analysis, and given that a vast majority of the messages have one underlying theme, we find it suitable to categorize them exclusively. Additionally, in non-experimental situations, we do not often consider people’s behavior to be “60% selfish, 20% apologetic, and 20% honest”; instead, we tend to generalize: The aforementioned person would usually be seen as selfish or even a jerk.

Furthermore, we require that the taker must play the game three times over the course of the experiment. This repetition allows us to determine whether takers’ behavior or rationale changes over time and, if it does, which differences in the scenario affect changes in their responses. And, because we slightly vary the rewards in each round of the experiment, this repetition causes takers to recalibrate their decisions. This prevents two rounds from ever being the same, and discourages takers from doing what they did in the last round. In sum, we conduct a three-round power-to-take game in which the assets start off with receivers and takers must explain their behavior when they select their take rates.

### B. Receiver experiment

Receivers and takers both completed the same set of questions before and after the power-to-take game part of the experiment, but receivers played against takers' data at a later date. Each receiver played against three takers. In each round, receivers were told their partner's partial MTurk ID, earnings for that round's effortful task, the amount their partner chose to take from them (the take rate), and their final earnings after the round had ended. They were also shown the message their partner sent explaining their actions unless, as a result of the taker's decisions, the explanation was not sent to the receiver. They, like the respondents who originally categorized takers' messages, were asked to categorize each of their partner's messages as apologetic, fair, guilty, honest, selfish, or not satisfactorily described by any of these adjectives. Then, receivers were asked on a five-point Likert Scale about the extent to which they agreed that they: thought their partner's message was sincere; wanted to play with that partner again; would pay part of their bonus to respond to their partner; would pay part of their bonus to see other messages sent by their partner; trust their partner; and were content with how their partner behaved and explained themselves. This process was repeated nine times, thrice for each of the three takers they played against.

### C. Treatments

We separate takers into the following twelve treatments to better investigate the different facets of their explanations and behavior. Participants did not change treatments during the experiment, and takers were assigned to only one treatment so that informational differences between the treatments did not affect the behavior of either party. The treatments are as follows:

*Informed* takers know before the experiment begins that they will be asked to explain their behavior later on. *Uninformed* takers do not have this information to help them plan an optimal set of future choices.

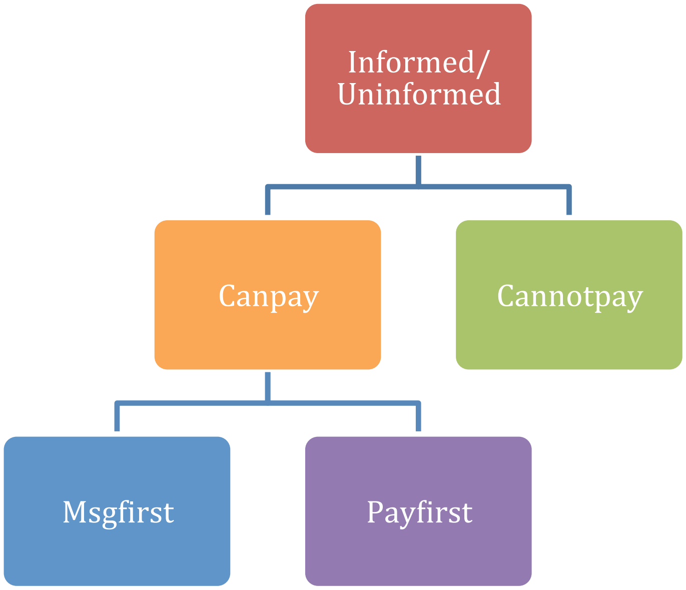
*Canpay* takers are given the option to select a wtp or wta to assuage the potentially uncomfortable explanation they find themselves forced to give. *Cannotpay* takers do not have this ability; their messages are always sent.

*Ascending* takers earn π in increasing order, {$0.20, $0.40, $0.80}, where each element is their profit in a given round, whereas *descending* takers earn π in decreasing order.

And finally, *msgfirst* takers, a subset that only applies to canpay takers, are asked to write a message to their partners before we elicit a wtp or wta from them; conversely, *payfirst* takers are asked if they want to pay before writing the message. We make this distinction to determine whether paying before or after writing a message affects decisions or the messages themselves.

Our overall experimental schematic is below. Note that the msgfirst, payfirst, and cannotpay nodes all have "ascending" and "descending" leaves not shown here.

Figure 1 – Treatment tree diagram



### D. Logistics

We ran our experiment on MTurk. No restrictions were placed on respondents except that they had to be located within the United States. We collected data from 587 participants over 3 rounds of the taker experiment, totaling 1761 observations, and from 195 individuals in the receiver experiment, totaling 1755 observations. The taker experiment took participants an average of 12 minutes to complete with average earnings of $1.41, whereas the receiver experiment took participants roughly 8 minutes with average earnings of $0.89. This rate of pay is commensurate with the MTurk average for high paying opportunities; more information about this and the MTurk sample population in general can be found elsewhere but is perhaps best summarized by Buhrmester et al. (2011).

### E. Hypotheses

With the aforementioned setup in place, we propose and test the following hypotheses:

First, we believe that *takers will have a positive willingness to pay* (or accept) for the transmission of their messages. In particular, we expect that takers, in an effort to look good, will select wtps (or wtas) to help facilitate the process of filtering out the ways in which receivers are able to receive their explanations. Along with this comes the assumption that explanations have value, which may be a result of guilt, regret, conflict aversion, fear of judgment or retribution, empathy, deceptive framing, or some other mechanism.

Second, we believe that *explanations will differ in cost*, which is to say that, since certain emotions are more painful to convey than others, different sorts of explanations will accompany different behaviors. Additionally, wtps (or wtas) should differ by type of message if their costs are truly different.

Third, we believe that *takers will treat the cost of explanations strategically* in the sense that they will seek to maximize their payoffs—recall that taker payoffs = $Y_T + tY_R + s(T, R)$—by estimating present and future values of explanations for both themselves and their receiver counterparts. They will shape their explanations to their behavior in such a way that maximizes this function. In particular, we suspect that the anticipated future cost of having to explain oneself will elicit more altruistic behavior, manifested in lower take rates, and less self-implicating explanations. Similarly, we suspect that takers with the ability to pay to avoid explanations will not see them as costly future events and will instead behave less altruistically, relying more on their explanations than their actions to maximize their profit functions. In other words, they use their explanations to counteract some of the negative emotions associated with explaining a higher take rate.

Fourth, we believe that *the act of explaining oneself is costly*. Accordingly, takers who fall in both the canpay and msgfirst treatments should have higher wtps than those in the payfirst treatments because the time and effort required to express one's emotions are costly. Additionally, takers may recognize that there are aspects of their explanations that they do not want to convey. Therefore they may pay more to prevent them from being sent, especially if they physically have to type out the message before sending it.

And fifth, we believe that *takers have imperfect foresight regarding the impact of their explanations*. In other words, though they try, they do not value their explanations correctly and therefore cannot plan their actions in a way that complements their explanations. This manifests itself in both experiments as takers' *ex ante* beliefs—i.e., that receivers will see their explanations in a given light—do not often align with receivers' *ex post* conceptions about them—that the way receivers feel about takers' explanations was the way takers wanted them to feel. This leads to receivers being unhappy with their taker counterparts and not wanting to maintain their relationships with them.

## III. Results from taker experiment

As we began to analyze the data, we realized that there were few important changes in behavior across rounds of the experiment. This is not to say that individual takers behaved in similar ways between rounds—as expected, they still recalibrated between rounds—but rather that takers' aggregate behavior did not change much over time. Accordingly, much of our analysis considers a panel data set where each participant in each round is treated as a unique observation. Though we retain round fixed effects in all of our regressions, they are not discussed at length since any significance in them can be attributed to takers refining their behavior over time or simply exploring more of their options as the experiment progresses.

### A. Summary statistics

Table 1 – Taker experiment summary statistics

| Variable | Mean | SD | Variable | Mean | SD |
|---|---|---|---|---|---|
| take (¢) | 44.1766 | 46.011 | age (years) | 31.8974 | 10.694 |
| takePos (%) | 0.5368 | 0.4988 | female (%) | 0.4141 | 0.4927 |
| wtp (¢) | 0.4704 | 2.4892 | apology (%) | 0.0313 | 0.1743 |
| wtpPos (%) | 0.0592 | 0.2362 | fairness (%) | 0.3675 | 0.4823 |
| wtpIfPos (¢) | 7.9403 | 6.7731 | guilt (%) | 0.0165 | 0.1275 |
| wta (¢) | 0.2095 | 1.2111 | honesty (%) | 0.1937 | 0.3953 |
| wtaPos (%) | 0.0575 | 0.2328 | selfishness (%) | 0.3322 | 0.4711 |
| wtaIfPos (¢) | 3.6462 | 3.629 | observations | 1755 | |

We first present summary statistics for our key variables of interest (above). A brief look at the overall distribution of takers' explanation types reveals that a majority of takers flocked to certain types of explanations: Those concerned with fairness and selfishness are more common, honest takers are not far behind, and only a minority of takers apologize or admit guilt.

Turning our attention to take rates, a full 46% of takers chose not to take any of their partners' earnings in the experiment, and those that did take took an average of only 44% of their partner's total earnings: This proportion of non-takers is high and this average take rate is low in comparison to similar power-to-take games (Ben-Shakhar et al., 2007; Grosskopf and Lopez-Vargas, 2014), likely because requiring that takers send explanations to their partners induces more altruistic behavior. This effect is reinforced by the actions of informed takers, which we will discuss shortly.

Overall, though most takers were either content with explaining their actions or preferred explaining themselves to relinquishing any of their earnings, several were not. Nearly 6% of canpay takers were willing to pay to try to prevent an explanation from being sent (wtpPos), and another 6% were willing to pay to try to guarantee that one was sent (wtaPos). Together, nearly 12% ($p < 0.01$, n=132) of the 1,131 canpay takers exhibited a willingness to exchange earnings for more control over their explanations, which is itself evidence of strategic behavior. It is also striking that the wtpPos takers on average chose to transfer 40% of the maximum they could, $0.20, to try to stop an explanation from being sent (wtpIfPos) while wtaPos takers chose to transfer only 18% of what they could (wtaIfPos). This difference suggests that takers felt fundamentally different about preventing explanations from being sent and guaranteeing that explanations were sent. This is reflected in both the idea that sending explanations—telling stories—has material value as well as in the distribution of messages within each subset of takers (see Table 2 below). WtpPos takers were significantly more selfish than wtaPos takers in their explanations, and it is therefore unsurprising that they were willing to pay more to avoid costlier explanations that would implicate themselves in greedy behavior.

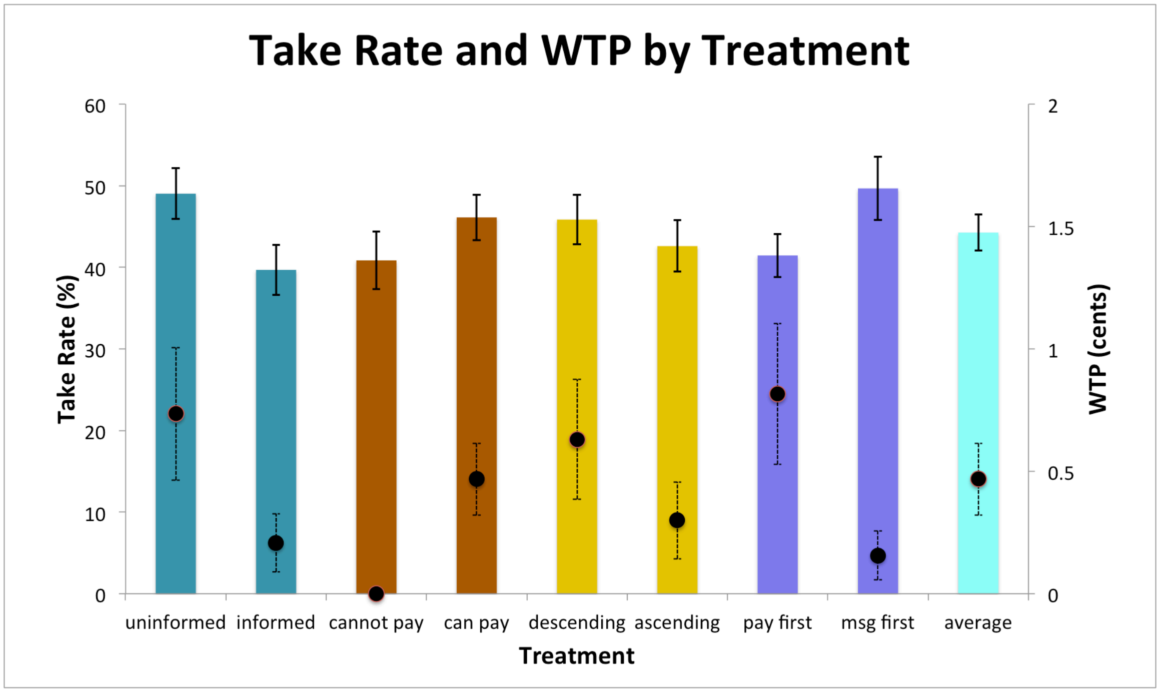
Table 2 – Frequencies of explanations for takers with positive wtp or wta

| Type | wtpPos | wtaPos |
|---|---|---|
| apology | 2 | 4 |
| fairness | 14 | 26 |
| guilt | 5 | 1 |
| honesty | 4 | 15 |
| selfishness | 34 | 14 |
| other | 8 | 5 |
| total | 67 | 65 |

We therefore find weak support for our first hypothesis and believe that the low rates of wtp/wta participation were likely due to profit maximizing strategies that valued the money saved in not paying to alter messages more heavily than the impact of the messages the payments could have altered. It is also worth noting that selecting a wtp/wta value below $0.20, the maximum we allowed, does not guarantee takers a certain result—recall that a random number is still drawn and compared to the wtp/wta value to govern the course of events—so risk averse takers may have preferred to grapple with an unpleasant explanation that they knew would transpire with complete certainty instead of a more damning explanation that may not have been sent with only partial certainty. The only way to control the fate of one’s messages with full certainty would have been to select a wtp/wta value of $0.20 in each round, an extremely costly choice that only two takers made (two others were wtp $0.20 in two of the three rounds). Predictably, the distribution of wtp values was different than that of wta values: a full 29 takers were wtp $0.10 or more, while 31 were wta $0.01, another 20 were wta $0.05, and only 6 were wta $0.10 or more. This analysis of takers’ decisions provides some amount of support for our second hypothesis, that messages differ in costliness, but to explore it more completely, we must first look at take rates and the differences between our treatments.

### B. Take rates and treatments

Figure 2 – Bar chart of take rates and wtps by treatment



In Figure 2, we find support for our third hypothesis, that people treat the costs of explanations strategically: As anticipated, informed takers behave significantly more altruistically than uninformed takers, and we also find significant differences in wtp between the two treatments. We attribute these differences to the timing of information takers have access to: By virtue of having more information upon which to base their original decisions, informed takers are comparatively willing to pay much less than they otherwise would be. It also seems that informed takers choose a lower-variance strategy than uninformed takers; by maintaining a lower take rate and wtp than uninformed takers, they reduce their exposure to any potential feedback from the receiver and generally try to minimize the receiver's effect on their payoffs. In our regression analysis (following page), we find that there are indeed strategic efforts behind takers' actions. Take rates play a significant role in determining the types of messages takers send to explain their actions, and the importance of informing select takers of the forthcoming explanations and granting them the ability to pay to avoid those explanations is further emphasized.

We also find that canpay takers tend to behave less altruistically than their cannotpay counterparts, but their wtp is not that high. This suggests that canpay takers deliberately plan a strategy around heavy reliance on explanations—and payments to control the transmission of those explanations—to mitigate costs associated with higher take rates; in other words, they accept the fact that they will try to explain or pay their way out of difficult situations. This possible payment effect is considerable but notably smaller than the effect information has on takers' decision-making processes. This may be both a case of both anchoring and timing: It is plausible that takers think of their interactions with receivers as a series of decisions. The first decision they make concerns take rate, and it is the most important since it governs the rest of their interactions with a given receiver. The second decision they make concerns explanations and, since it is conditional on the original take rate decision, it is less important to takers and hence has smaller effects on their behavior. Finally, the third decision takers make concerns paying to avoid explanations, which is another step removed from the original take rate decision and therefore inferior still. These diminishing returns to importance with respect to interactions lend support to our fifth hypothesis, that takers have imperfect foresight regarding their explanations as they relate to their actions.

Table 3 – OLS regressions with take rate as dependent variable

| VARIABLES | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| age | -0.563*** | -0.563*** | -0.157 | -0.177* | -0.182* |
| | (0.183) | (0.184) | (0.0994) | (0.0962) | (0.0947) |
| female | -12.13*** | -12.13*** | -2.835 | -2.863 | -3.270 |
| | (3.711) | (3.714) | (2.029) | (2.045) | (2.034) |
| agreeableness | -10.83*** | -10.83*** | -3.053*** | -2.956*** | -2.769*** |
| | (1.816) | (1.817) | (1.031) | (1.023) | (0.964) |
| informed | -9.425*** | -9.425*** | -4.608** | -8.436* | -2.983 |
| | (3.607) | (3.610) | (1.939) | (4.808) | (10.16) |
| canpay | -1.842 | -1.842 | 0.964 | -10.28** | -1.729 |
| | (4.342) | (4.344) | (2.340) | (4.762) | (12.58) |
| ascending | -2.286 | -2.286 | -2.598 | -10.86** | -17.88* |
| | (3.529) | (3.531) | (1.838) | (4.500) | (10.07) |
| msgfirst | 8.098* | 8.098* | 2.678 | 6.942 | 30.29** |
| | (4.535) | (4.538) | (2.351) | (4.903) | (12.78) |
| round_2 | | 5.470*** | 3.188** | 3.203** | 3.235*** |
| | | (1.096) | (1.259) | (1.257) | (1.249) |
| round_3 | | 7.113*** | 3.048** | 3.080** | 3.260** |
| | | (1.300) | (1.337) | (1.337) | (1.325) |
| apology | | | 41.94*** | 41.43*** | 54.14*** |
| | | | (5.819) | (5.813) | (14.19) |
| fairness | | | -36.22*** | -36.07*** | -14.84 |
| | | | (5.063) | (5.040) | (10.12) |
| guilt | | | -3.732 | -4.668 | -9.235 |
| | | | (9.641) | (9.744) | (20.62) |
| honesty | | | -40.16*** | -40.04*** | -26.71** |
| | | | (5.191) | (5.165) | (10.58) |
| selfishness | | | 37.76*** | 37.38*** | 46.73*** |
| | | | (5.054) | (5.033) | (10.24) |
| constant | 92.54*** | 88.34*** | 63.16*** | 69.81*** | 57.24*** |
| | (15.99) | (16.03) | (10.20) | (10.47) | (13.36) |
| round fx | No | Yes | Yes | Yes | Yes |
| explanations as indep vars | No | No | Yes | Yes | Yes |
| treatmentXtreatment intfx | No | No | No | Yes | Yes |
| treatmentXmessage intfx | No | No | No | No | Yes |
| Observations | 1,626 | 1,626 | 1,626 | 1,626 | 1,626 |
| R-squared | 0.112 | 0.116 | 0.653 | 0.657 | 0.674 |

Robust standard errors in parentheses: *** p<0.01, ** p<0.05, * p<0.1

We investigate our fourth hypothesis, that act of physically explaining oneself is costly, by comparing the actions of payfirst and msgfirst takers, who were both subsets of canpay takers, to the standard of canpay takers in general. Our expectation was that msgfirst takers would have higher wtps than payfirst takers because of the time and effort costs associated with creating an explanation. We actually find the contrary (see Table 3 above and Table 4 below): Msgfirst takers took more and were wtp less than the general canpay population, while payfirst takers took less and were wtp more. Table 4 also provides support for our second hypothesis that messages differ in costliness; the largest effects on wtp are associated with fairness and honesty messages, which people are more inclined to send because doing so does not implicate them in misbehavior.

Table 4 – OLS regressions with wtp as dependent variable

| VARIABLES | (1) | (2) | (3) |
|---|---|---|---|
| informed | -0.391* | -0.391* | -0.422* |
| | (0.223) | (0.223) | (0.215) |
| msgfirst | -0.660*** | -0.660*** | -0.692*** |
| | (0.244) | (0.244) | (0.239) |
| apology | | | -0.527 |
| | | | (0.576) |
| fairness | | | -0.879** |
| | | | (0.422) |
| guilt | | | 1.810 |
| | | | (1.182) |
| honesty | | | -1.105** |
| | | | (0.476) |
| selfishness | | | -0.605 |
| | | | (0.478) |
| round fx | No | Yes | Yes |
| explanations as indep vars | No | No | Yes |
| Observations | 1,053 | 1,053 | 1,053 |
| R-squared | 0.082 | 0.082 | 0.113 |

Robust standard errors in parentheses: *** p<0.01, ** p<0.05, * p<0.1

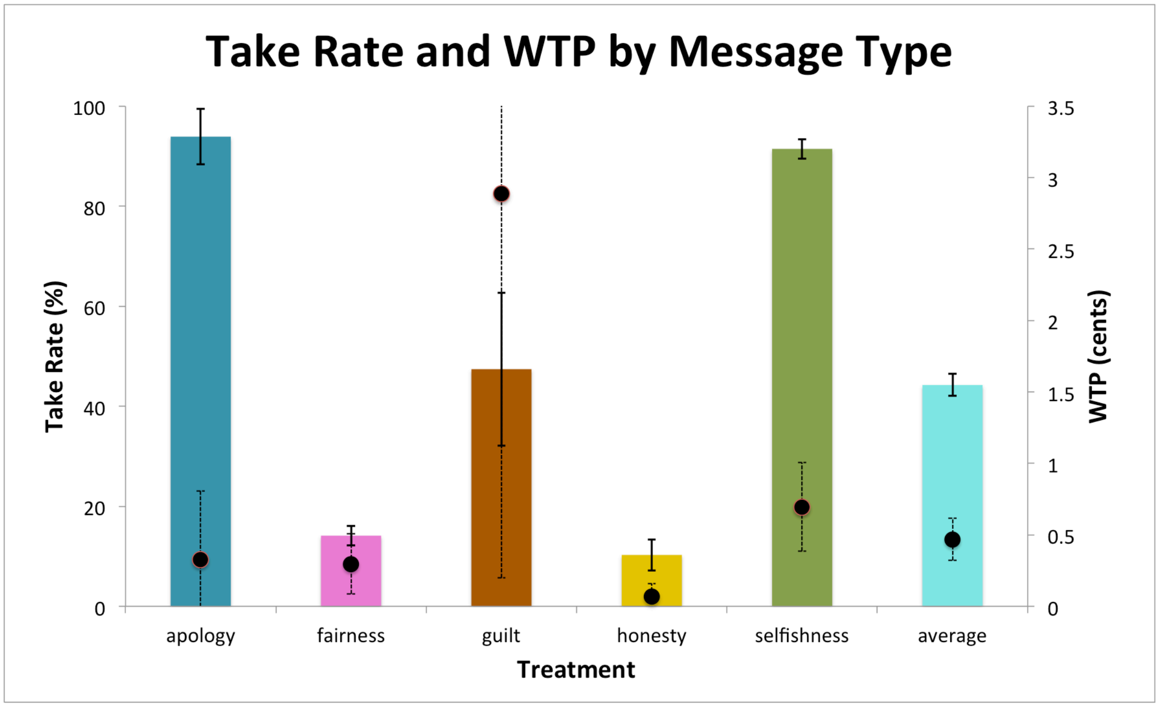
In a way, msgfirst takers acted as if they were uninformed takers (high take rate) but were actually thinking with an aggressive form of informed logic (low wtp, low reliance on explanation). We see this in the types of messages they send: msgfirst takers, the more aggressive of the two flavors of canpay takers, were nearly 30% more likely to apologize, 20% more likely to admit selfishness, and 25% less likely to be honest than payfirst takers. In contrast, payfirst takers were relatively meek; they did not take very much and were relatively averse to their messages being sent. We can also investigate message costliness for informed and canpay treatments to provide further support for our second hypothesis. Most of Table 5 (below) is unsurprising: Informed takers' explanations were more concerned with fairness and honesty than selfishness, whereas canpay takers' were less fair and more selfish. The surprising result is that associated with apologetic takers who seemed to employ a different strategy altogether.

Table 5 – Explanation frequencies in informed and canpay treatments

| Explanation type | Uninformed | Informed | Cannotpay | Canpay |
|---|---:|---:|---:|---:|
| apology | 0.47% | 5.69% | 1.44% | 4.07% |
| fairness | 34.97% | 38.46% | 39.42% | 35.28% |
| guilt | 1.40% | 1.90% | 1.76% | 1.59% |
| honesty | 18.41% | 20.29% | 18.43% | 19.89% |
| selfishness | 39.86% | 26.87% | 32.21% | 33.78% |
| other | 4.90% | 6.80% | 6.73% | 5.39% |

## C. Costliness of messages

Figure 3 – Bar chart of take rates and wtps by message type



To investigate this strategy further, we look at take rates and wtps by message type. Fairness, guilt, honesty, and selfishness takers all behave as one might expect, but the most striking aspect of this chart is the case of informed takers, who behave even more selfishly than self-proclaimed selfish takers. This is even more surprising given that most of the apologetic takers were in informed treatments, which we have thus far associated with “good” behavior. However, we have been overlooking a fierce strategy that only these takers employ: By taking extremely high amounts—51 of the 55 apologetic takers were informed, and only 5 of the 55 had take rates < 100%—selecting low wtps, and giving explanations that people generally think of as being compassionate, apologetic takers obtained very high payoffs with very low emotional risk because of the social effects of apologies. This "misuse" of apologies contradicts much of the extant literature (Ho 2012) that suggests apologies are used to build and strengthen relationships. We postulate that information timing brings about this tendency. By learning ahead of time that they will need to explain themselves later, informed takers have the ability to select an explanation that makes them seem less self-centered in the eyes of receivers. If these takers believe that the cost-mitigating aspect of making an apology outweighs any amount of monetary wrongdoing that they can do to receivers, then it makes sense for them to choose this strategy in an effort to maximize their wellbeing.

## IV. Results from receiver experiment

This brings us to our fifth hypothesis, that people have imperfect foresight about the impact of their explanations, and we investigate this in the receiver experiment. First we reiterate the assumption that if takers are properly valuing their explanations, then receivers should receive their explanations in the ways that takers intended for them to be received; in other words, selfishness messages should come across as selfish, apologies should be apologetic, etc. If this is not the case, it is unlikely that takers can adequately plan their actions in a way that complements their explanations, and they may be making misconceptions about how their behavior is actually received. A better understanding of their explanations' reception would help them make more profitable decisions in the future.

Recall that the receiver experiment was structured around several questions relating to receivers' satisfaction with and perceptions about their taker counterparts. Instructions and regressions for the second experiment are available in the appendix.

### A. Summary statistics

Table 6 – Receiver experiment summary statistics

| Variable | Mean | SD |
|---|---|---|
| contentedness | 2.4931 | 1.5638 |
| sincerity | 3.1258 | 1.1341 |
| trustworthiness | 2.1684 | 1.5746 |
| seeOtherMsgs | 0.7861 | 1.1614 |
| playAgain | 2.1586 | 1.7352 |
| payToRespond | 0.9742 | 1.2757 |
| observations | 1755 | |

We present summary statistics for the six main variables in the receiver experiment in Table 6. All variables were measured on five-point Likert scales. The seventh question asked takers to characterize the explanations takers sent them and will be discussed shortly.

### B. Impact of treatments

The considerable treatment effects found in the taker experiment are largely absent from the receiver experiment. The most significant finding is that receivers partnered with informed takers were about 12% more content with their behavior, 16% more interested in playing with them in subsequent rounds, and 14% more trusting of them than they were of uninformed takers. This suggests that, since most informed takers behaved altruistically, they were successful in framing their actions in ways that were appealing to receivers.

These, and subsequent results, are summarized in Table 7 below. Note that we dropped take rate from all regressions because of endogeneity concerns. Independent variables are listed as column headings.

Table 7 – OLS regressions of receiver responses to taker behavior (on a 1-5 scale)

| VARIABLES | contentedness | sincerity | trustworthiness | seeOtherMsgs | playAgain | payToRespond |
|---|---|---|---|---|---|---|
| female | 0.106 | 0.228*** | 0.149** | 0.157** | 0.138** | 0.190*** |
| | (0.0683) | (0.0608) | (0.0657) | (0.0635) | (0.0622) | (0.0678) |
| agree | 0.129*** | 0.0271 | 0.132*** | -0.0442 | 0.152*** | -0.0331 |
| | (0.0352) | (0.0314) | (0.0340) | (0.0328) | (0.0320) | (0.0350) |
| informed | 0.150** | 0.00138 | 0.189*** | -0.00631 | 0.188*** | -0.0298 |
| | (0.0652) | (0.0581) | (0.0629) | (0.0607) | (0.0594) | (0.0647) |
| ascending | 0.0813 | 0.192*** | 0.0269 | 0.0487 | 0.119** | 0.0243 |
| | (0.0634) | (0.0564) | (0.0611) | (0.0590) | (0.0576) | (0.0629) |
| canpay | -0.0457 | -0.123* | -0.00577 | 0.127* | -0.0476 | 0.0503 |
| | (0.0789) | (0.0701) | (0.0759) | (0.0734) | (0.0716) | (0.0783) |
| msgfirst | 0.00407 | 0.130* | -0.0813 | -0.0411 | 0.0152 | -0.0271 |
| | (0.0792) | (0.0705) | (0.0763) | (0.0736) | (0.0719) | (0.0787) |
| apology | -0.996*** | -0.552*** | -1.001*** | -0.460** | -1.679*** | -0.783*** |
| | (0.214) | (0.189) | (0.204) | (0.198) | (0.193) | (0.212) |
| fairness | 1.017*** | 0.561*** | 1.284*** | 0.0652 | 1.358*** | 0.172 |
| | (0.142) | (0.126) | (0.136) | (0.132) | (0.129) | (0.141) |
| guilt | -0.267 | -0.168 | -0.0654 | 0.513** | -0.0308 | 0.333 |
| | (0.257) | (0.229) | (0.248) | (0.243) | (0.234) | (0.256) |
| honesty | 0.930*** | 0.519*** | 1.191*** | 0.171 | 1.243*** | 0.176 |
| | (0.150) | (0.133) | (0.143) | (0.139) | (0.136) | (0.149) |
| selfishness | -1.091*** | 0.0381 | -0.942*** | -0.432*** | -1.441*** | -0.607*** |
| | (0.143) | (0.127) | (0.137) | (0.133) | (0.130) | (0.142) |
| msglength | 0.00112* | 0.00161*** | 0.00144** | 0.00110** | 0.00103* | 0.000794 |
| | (0.000583) | (0.000519) | (0.000561) | (0.000544) | (0.000530) | (0.000580) |
| round_2 | -0.122 | -0.122* | -0.112 | -0.0786 | -0.146** | -0.107 |
| | (0.0767) | (0.0685) | (0.0741) | (0.0715) | (0.0699) | (0.0763) |
| round_3 | -0.104 | -0.153** | -0.0607 | -0.0423 | -0.138** | -0.0729 |
| | (0.0774) | (0.0688) | (0.0744) | (0.0720) | (0.0702) | (0.0768) |
| Constant | 2.064*** | 1.590*** | 1.799*** | 0.794** | 2.001*** | 1.096*** |
| | (0.336) | (0.299) | (0.323) | (0.313) | (0.306) | (0.334) |
| Observations | 1,443 | 1,454 | 1,448 | 1,448 | 1,448 | 1,448 |
| R-squared | 0.440 | 0.142 | 0.489 | 0.127 | 0.625 | 0.153 |

Standard errors in parentheses: *** p<0.01, ** p<0.05, * p<0.1

## C. Impact of explanations

Much like in the taker experiment, the different kinds of messages takers used to explain themselves had a large impact on their behavior and consequently their reception by receivers. Receivers who received fairness and honesty messages were very much pleased with their partners' behavior, whereas those who received apology and selfishness messages were not. Reception of apologetic messages was particularly poor and even surpassed selfishness messages in eliciting negative responses from receivers. It is also worth noting that receivers were less interested in seeing the other messages takers sent and paying to respond to them when they had been wronged by behavior associated with a greedier message. This effect is absent for fairness, guilt, and honesty messages and suggests that receivers who had been wronged were mostly concerned with how they, themselves, had been treated and not how other people in general fared. Thus the taker strategy of sending greedy messages—which are associated with greedy actions and cause negative emotions on the receivers' end—leads to a lack of pro-social behavior. Lastly, we found that recipients of longer messages were more pleased with their partners as well.

## D. Impact of explanation classifications

Table 8 – Comparison of message categorization between experiments

| Taker classification | Receiver classification | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | apology | percent | fairness | percent | guilt | percent | honesty | percent | selfishness | percent |
| apology | 11 | 20.0% | 0 | 0.0% | 4 | 7.3% | 17 | 30.9% | 22 | 40.0% |
| fairness | 8 | 1.2% | 427 | 66.4% | 22 | 3.4% | 150 | 23.3% | 21 | 3.3% |
| guilt | 7 | 24.1% | 4 | 13.8% | 7 | 24.1% | 8 | 27.6% | 3 | 10.3% |
| honesty | 6 | 1.8% | 171 | 50.3% | 20 | 5.9% | 97 | 28.5% | 14 | 4.1% |
| selfishness | 17 | 2.9% | 17 | 2.9% | 17 | 2.9% | 181 | 31.2% | 329 | 56.6% |

We find that many messages were categorized differently in the second experiment than in the first (see Table 8, above). For example, only 20% of messages categorized as apologies in the first experiment were received as apologies in the second; twice as many were considered selfish, a finding that summarizes how receivers felt about apologetic takers' behavior. Fairness messages held up well, and many honesty messages were received as fair, likely because the categories are closely related. Despite the small sample size, guilt messages were received in mixed ways, perhaps suggesting either that people do a relatively poor job of signaling guilt when they mean it or a relatively good job of hiding it when they do not. Lastly, a large number of selfishness messages were seen as honest. Overall, this finding supports our fifth hypothesis since people do not fully understand how they are actually presenting themselves to others. Thus they often employ a suboptimal strategy predicated on misunderstanding or use deceptive framing to avoid saying what they actually mean for strategic purposes.

It is also worth noting that there were no significant gender effects present in receivers' classification of messages. While men were more likely to send greedy messages, neither men nor women were more likely to receive them, or any other kinds of message, differently.

### E. Impact of gender

Perhaps the most interesting and socially significant finding of the second experiment was that, even after controlling for all relevant variables relating to both takers and receivers, all subsets of receivers significantly preferred to interact with female takers rather than male ones (refer to Table 7). The magnitude of the effect was comparable to that of being partnered with an informed taker, but it was prevalent in more of the questions asked of receivers; it was significant in the contentedness question at the 12% significance level, the only question above 10%. This suggests that women inherently express themselves differently than men do in ways that we have not accounted for. And, short of conducting detailed text analysis on the explanations takers sent, we conclude that the reasoning for this effect is beyond the initial scope of our experiment. Accordingly, it is dangerous to over-interpret the result, but we encourage research efforts directed towards that and related questions.

## V. Discussion

In the introduction, we set out to untangle the strategy behind human decision-making in an effort to better understand the economic value of costly explanations. We claimed that takers would treat their explanations strategically and found support for this in the fact that their behavior changed drastically between treatments. In particular, we note that information, and particularly the timing of that information, plays a substantial role in shaping takers' decisions. From a theoretical perspective, both informed and canpay takers obtain knowledge about the future that they can use to make better decisions in the present, yet the information comes framed in different ways. Takers in informed treatments learn ahead of time about a certain future explanation, which may be seen as a cost by some, whereas canpay takers are informed ahead of time about an entirely optional future cost. As such, we are offering the different subsets of takers information about their own "wellbeing futures" and letting them set the price. Additionally, there may also exist a physiological timing difference between the two: It may be the case that humans make decisions that trigger our informed brains (which answer the question “How should I act given this future scenario?”) before triggering our canpay brains (which answer “How do I reconcile my behavior with this explanation?”), but again, investigating this question is beyond the scope of our experiment.

Nonetheless, when told ahead of time about future explanations, informed takers behaved more altruistically and explained themselves in ways that better met receivers' expectations. Contrariwise, when told beforehand that they would be able to pay to avoid future explanations, canpay takers behaved less altruistically and put more of their actions' weight on their explanations. In other words, they acted selfishly, tried to explain their way out of upsetting their partners, and were willing to pay more to try to do so. This supports another of our hypotheses, that takers have a positive willingness to pay (or accept) for control over their explanations' transmission. We also explain a time-dependent mechanism for why only an eighth of takers were willing to exchange earnings for a control over the receipt of their explanations: As takers make repeated decisions in a given interactions, subsequent decisions become increasingly less valuable to them. Thus, by the time they are considering whether to select a positive wtp/wta value—no less than their third decision along the way—they have already made more important decisions, so this one does not matter as much in comparison.

We also argued that explanations differ in cost and find support for this in the fact that different treatments were very strongly associated with sending different types of explanations and accordingly, with vastly different strategic approaches to the game. Willingnesses to pay were also highly disparate between treatments. Associated with this argument was the claim that the act of explaining oneself is in itself costly, for which we find support for the opposite; takers who sent their messages first were less altruistic and less willing to pay than those who paid before sending messages, which supports the ideas that people are generally interested in expressing themselves and the costs associated with doing so are small.

Finally, we argued that takers have imperfect foresight about the impact of their explanations, and we find support for this as well. Apologies, our best example of deceptive framing, were almost negligibly costly for takers (even less so than self-implicating selfish explanations), but the fact that receivers did not receive them well suggests that they were misused as a result of takers misunderstanding their strategic value. Fairness and honesty messages, on the other hand, were low-cost, low-risk alternatives that generally sat well with receivers.

## V. Conclusion

As such, these results paint an interesting picture of the taker role in society and the strategy of explanations in general. While it is clear that we attempt to strategize when it comes to explaining ourselves, it is equally clear that we do not do so very well. It is likely that we overvalue our own payoffs—i.e., rely too heavily on mechanisms such as deceptive framing—and do not consider the effects that our explanations have on others—show empathy—nearly enough. And, while it is not always our intent to deceive others, it often happens that we do so as a result of failing to understand how our actions affect them. This makes our second experiment all the more relevant. Each of the questions posed in the receiver experiment dealt with a social practice: The contentedness, sincerity, trustworthiness, and playAgain questions had to do with different aspects building and maintaining relationships; the seeOtherMsgs question had to do with social "snooping"; and the payToRespond question had to do with retribution. These are all pertinent question in the study of interpersonal interactions, and the fact that female takers fared better than male ones in all of them suggests that women are inherently better at explaining themselves in ways that appeal to receivers, which is significant. This finding would imply that, women should more readily make costly explanations than men. And, though this is perhaps an overstatement of the findings, it is clear that these results carry weight in studying matters of politics, finance, and other fields that deal with interpersonal interactions.

But in some ways, the most relevant treatment in our experiment is the uninformed, canpay one since it is the “default” state of most people in everyday life: They are not actively thinking about their futures but do have some amount of agency over the course of their interactions. Per our findings, these people are prone to low levels of altruistic behavior, self-serving explanations, and, in general, little concern for other people. This seems like a fairly bleak perspective, but, by informing people about their futures, it may be possible to shift them into the informed, canpay condition where they are more likely to engage in altruistic and pro-social behavior and send well-received explanations. Many would argue that this is more desirable state of existence, one in which we recognize the value of our explanations and can more easily behave optimally in interpersonal interactions. Thus there is need for more research on the topic of the economics of stories. There is a wealth of possible applications for modeling behavior based on communication and inducing changes in behavior in a predictable fashion. The fields of politics and finance immediately come to mind: Imagine how would the world be different if we could more easily encourage altruistic behavior in politicians, persuade world leaders to behave more pro-socially, or even simply detect lies and false apologies more accurately.

## VI. References

Andreoni, James, and Justin M. Rao. "The power of asking: How communication affects selfishness, empathy, and altruism." *Journal of Public Economics* 95, no. 7 (2011): 513-520.

Batson, C. Daniel, Janine L. Dyck, J. Randall Brandt, Judy G. Batson, Anne L. Powell, M. Rosalie McMaster, and Cari Griffitt. "Five studies testing two new egoistic alternatives to the empathy-altruism hypothesis." *Journal of personality and social psychology* 55, no. 1 (1988): 52.

Ben-Shakhar, Gershon, Gary Bornstein, Astrid Hopfensitz, and Frans van Winden. "Reciprocity and emotions in bargaining using physiological and self-report measures." *Journal of Economic Psychology* 28, no. 3 (2007): 314-323.

Bosman, Ronald, and Frans Van Winden. "Emotional Hazard in a Power-to-Take Experiment." *The Economic Journal* 112, no. 476 (2002): 147-169.

Bosman, Ronald, and Frans Van Winden. "The behavioral impact of emotions in a power-to-take game: An experimental study." (2000).

Bosman, Ronald, Matthias Sutter, and Frans van Winden. "The impact of real effort and emotions in the power-to-take game." *Journal of Economic Psychology* 26, no. 3 (2005): 407-429.

Buhrmester, Michael, Tracy Kwang, and Samuel D. Gosling. "Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data?" *Perspectives on psychological science* 6, no. 1 (2011): 3-5.

Charness, Gary, and Martin Dufwenberg. "Promises and partnership." *Econometrica* 74, no. 6 (2006): 1579-1601.

Dufwenberg, Martin, and Uri Gneezy. "Measuring beliefs in an experimental lost wallet game." *Games and Economic Behavior* 30, no. 2 (2000): 163-182.

Ellingsen, Tore, and Magnus Johannesson. "Anticipated verbal feedback induces altruistic behavior." *Evolution and Human Behavior* 29, no. 2 (2008): 100-105.

Ho, Benjamin. "Apologies as signals: with evidence from a trust game." *Management Science* 58, no. 1 (2012): 141-158.

Galeotti, Fabio. On the Robustness of Emotions and Behavior in a Power-to-Take Game Experiment. No. 13-07. School of Economics, University of East Anglia, Norwich, UK., 2013.

Grosskopf, Brit, and Kristian Lopez-Vargas. "On The Demand for Expressing Emotions." (2014).

Kahneman, Daniel, Jack L. Knetsch, and Richard H. Thaler. "Fairness and the assumptions of economics." *Journal of business* (1986): S285-S300.

List, John A. "On the interpretation of giving in dictator games." *Journal of Political Economy* 115.3 (2007): 482-493.

Reuben, Ernesto, and Frans Van Winden. *Negative reciprocity and the interaction of emotions and fairness norms*. No. 1685. CESifo working papers, 2006.

Vanberg, Christoph. "Why do people keep their promises? An experimental test of two explanations." *Econometrica* 76, no. 6 (2008): 1467-1480.

Xiao, Erte, and Daniel Houser. "Emotion expression and fairness in economic exchange." *University of Pennsylvania* (2007).

Xiao, Erte, and Daniel Houser. "Emotion expression in human punishment behavior." *Proceedings of the National Academy of Sciences of the United States of America* 102, no. 20 (2005): 7398-7401.

# VII. Appendix

## A. Taker experiment instructions

Informed, canpay, ascending, msgfirst: http://goo.gl/3lZ3ay

Informed, canpay, ascending, payfirst: http://goo.gl/Dk5KAT

Informed, canpay, descending, msgfirst: http://goo.gl/l2EMm7

Informed, canpay, descending, payfirst: http://goo.gl/txT9YN

Uninformed, canpay, ascending, msgfirst: http://goo.gl/OjQ2qH

Uninformed, canpay, ascending, payfirst: http://goo.gl/GJE2u5

Uninformed, canpay, descending, msgfirst: http://goo.gl/UNfw2C

Uninformed, canpay, ascending, payfirst: http://goo.gl/JWGGBP

Informed, cannotpay, ascending: http://goo.gl/WFMYty

Informed, cannotpay, descending: http://goo.gl/VdKq6c

Uninformed, cannotpay, ascending: http://goo.gl/cXzsrC

Uninformed, cannotpay, descending: http://goo.gl/6RalWR

## B. Receiver experiment instructions

All receiver treatments: http://goo.gl/tnWBwr