What You Don’t Know Won’t Hurt You: A Laboratory Analysis of Betrayal Aversion

Jason A. Aimone¹ & Daniel Houser²

Interdisciplinary Center for Economic Science
George Mason University

August, 2010

Abstract: Recent research argues “betrayal aversion” leads many people to avoid risk more when a person, rather than nature, determines the outcome of uncertainty. Unfortunately, previous studies conflate betrayal aversion with established preference effects including loss-aversion. Using a novel investment-game experiment that varies how strategic uncertainty is resolved, we here provide rigorous evidence on detrimental effects of betrayal aversion. The impact is substantial: holding fixed the probability of betrayal, the possibility of knowing that one has been betrayed reduces investment by about one-third. We suggest emotion-regulation underlies this result and explains the importance of impersonal, institution-mediated exchange in promoting economic efficiency.

JEL Classifications: C91; D03; D81

Keywords: Betrayal Aversion, Risk, Trust, Emotion Regulation
What You Don’t Know Won’t Hurt You:
A Laboratory Analysis of Betrayal Aversion

August, 2010

Abstract: Recent research argues “betrayal aversion” leads many people to avoid risk more when a person, rather than nature, determines the outcome of uncertainty. Unfortunately, previous studies conflate betrayal aversion with established preference effects including loss-aversion. Using a novel investment-game experiment that varies how strategic uncertainty is resolved, we here provide rigorous evidence on detrimental effects of betrayal aversion. The impact is substantial: holding fixed the probability of betrayal, the possibility of knowing that one has been betrayed reduces investment by about one-third. We suggest emotion-regulation underlies this result and explains the importance of impersonal, institution-mediated exchange in promoting economic efficiency.
1 Introduction

First movers in bargaining environments often must decide whether to forgo guaranteed returns in order to trust counterparts to provide them with greater future benefits (e.g., efficiency wages in labor markets; see Fehr and Falk 1999; Rigdon, 2002). While experiments in both naturally occurring and designed environments have established that many people do trust, some people of course do not (see, e.g., Berg et al. 1995). Choosing not to trust is consistent with standard economic theory, and perhaps for this reason few alternative explanations for such decisions have appeared. This is unfortunate both because choosing not to trust might involve decision processes distinct from those that standard theory suggests, and because designing institutions to promote trust requires one to understand why sometimes people choose not to do so.

To explain the decision not to trust, standard theory posits that selfish people consider the stochastic nature of their environment, forming subjective estimates of the likelihood of reciprocity, and then make the expected earnings-maximizing decision. We provide evidence from a laboratory analysis of an investment game that this explanation is incomplete. In particular, we find that people’s trusting decisions vary according to how an environment’s uncertainty is resolved. If resolving uncertainty requires an investor to learn whether her trustee chose to betray her then she is much less likely to trust. Accordingly, our data provide evidence that “betrayal aversion” detrimentally affects propensities for trusting decisions, thus emphasizing the importance of impersonal, institution-mediated exchange in promoting trust, investment and efficient economic outcomes.

Recent interest in effects of betrayal aversion on trust decisions owes largely to research by Bohnet and Zeckhauser (2004). That paper, and the papers that follow it such as Hong and Bohnet (2007) and Bohnet et al. (2008), use the same experiment design. In brief, the investigators asked each subject to report the “minimum acceptable probability” (MAP) at which s/he would choose a (trust or risk) “gamble.” Consistent with betrayal aversion and the previous research on the distinction between risk and trust games (Houser et al. 2008, Eckel and Wilson 2004, etc.), these papers found subjects report higher average MAPs for trust gambles than risk gambles. However, as noted by Bohnet and Zeckhauser (2004), several other factors unrelated to betrayal could be at work in generating differences in behavior between treatments. They point to disutility from loss of control, assessment costs associated with calculating trustworthiness,
costs of making incorrect assessments, costs from placing trustees in a potentially undesirable decision situation, and disutility from earning money due to other people’s kindness as factors that could lead to differences between treatments.

For example, due to potential differences in reference points between MAP treatments, loss aversion may drive treatment differences. Indeed, Bohnet et al. (2006) discuss data from this same design within a loss aversion framework that does not appeal to betrayal considerations. Further, recent research (Trhal and Radermacher, 2009) shows that altruism concerns are heavily influenced by the decisions of the person on the receiving end of the altruistic act. These findings raise the concern that decreased willingness to accept risk (higher MAPs) in the trust games (compared to the MAPs in the risk games) may be due to differences in one’s counterpart’s roles between treatments: counterparts are passive in the risk game but active in the trust game. In light of this, whether results from a MAP design can be traced to betrayal aversion is unclear.

Here we report data from one-shot two-person binary investment games (Tullock, 1967) in which investors can choose not to know the decision of their particular trustee, and instead receive payment according to a random draw from a separate pool of decisions identical to the pool of trustees’ decisions. Note that the probability of receiving the “cooperative” outcome is identical in the two cases, and participants understand this is the case. Our design differs from the MAP-based design primarily in that it does not require one to elicit probabilities: our inferences stem from revealed-preferences. We argue below that this is a significant improvement in the sense that it allows us to avoid certain confounds that make the interpretation of MAP-based designs difficult. Specifically, our experiment design holds the position of trustees constant between treatments, removing loss-aversion and altruism concerns.

Our main finding is that, due to betrayal aversion, investors systematically prefer to remain ignorant of their specific trustee’s decision. Moreover, when avoiding this information is not possible investors are substantially less likely to make trusting decisions. These results are convergent evidence that outcome-based models cannot fully explain economic decision making in strategic environments (see, e.g., McCabe, et al. 2003). More specifically, our findings suggest that impersonal institution-mediated exchange (e.g., lending through financial intermediaries)

---

3 Tullock (1967) is the first paper, to our knowledge, to discuss and work through the implications of a sequential prisoner’s dilemma within a trust context.
promotes trusting decisions and economic efficiency by shielding investors from knowing whether their particular trustee chose to betray.

Our data also reinforce the general importance of emotions to economic decision-making (see, e.g., Fehr et al. 2005; Xiao and Houser, 2005), and provide new evidence to the importance of emotion regulation (see, e.g., Gross, 1998; Miu et. al 2008). Moreover, our design suggests a way to construct institutional solutions to inefficiencies stemming from betrayal aversion, an important topic that previous investigations have not addressed. Additionally, our experiments demonstrate that the majority of people choose such institutions when provided institutional choice, suggesting the importance of designing institutions with emotion regulation in mind.

Finally, our investigation contributes to the literature on distinctions between trust and risk environments (see, e.g., Houser et al., 2010; Schechter, 2007; Kosfeld et al., 2005; Eckel and Wilson, 2004; Ashraf et al. 2006; McCabe et al., 2001; Snijders and Keren, 1998). This literature provides substantial evidence that people make trusting decisions differently than decisions under risk. However, identifying the source of the differences is difficult. A reason is that trust and risk environments typically differ in multiple ways (e.g., strategic-uncertainty always involves another person, while state-uncertainty need not.) As a result, while it is widely accepted that trusting decisions differ between environments of strategic and state-uncertainty, the reason for such differences – and particularly the role of betrayal aversion – remains an important open question.

The following section reviews the relevant betrayal literature. Our experiment design is in section 3. Section 4 gives predictions and hypotheses, and section 5 describes our data and results. Section 6 is a concluding discussion.

2 Background

To experience betrayal is exceptionally emotionally costly. As noted by Jackson (2003), “the betrayed experience powerful sensations of violation; they feel used and damaged” (see also Finkel et al., 2002). Humans respond strongly to the thought of betrayal, whether betrayal is the result of a violation of one-sided trust (such as a friend’s non-repayment of a loan) or of mutual

---

4 In our case, emotion regulation can explain the actions people take to avoid the negative emotional experience of betrayal.

5 Jackson, p. 72.
trust (such as infidelity in a marriage.) These strong responses occur even when the betrayal does not involve them. For example, survey evidence gathered by Koehler and Gershocoff (2003) indicates that people punish hypothetical crimes including an aspect of betrayal much more severely than similar hypothetical crimes lacking this element.

Humans surely prefer not to experience betrayal and, consequently, might prefer to avoid situations where betrayal might occur. Such avoidance decisions are consistent with the notion of “emotion-regulation” (ER) which has emerged in recent psychology and neuroscience literature. Studies in this area (see, e.g., surveys by Gross (1998) and Ochsner and Gross (2005)) argue that ER, at both the conscious and subconscious level, is an important reason that expectations affect decisions. For example, it can explain why people often choose to turn conversations away from topics they expect could become uncomfortable. More generally, it can underlie any decision not to take an action that one expects might lead to unpleasant emotional outcomes. Indeed, ER could explain behavior observed in several recent studies where people seek to avoid guilt or shame by, for example, refraining from making promises (Charness and Dufwenberg, 2006) or avoiding freely accessible information (Dana, Cain, and Dawes, 2006; Dana, Weber and Kuang, 2007).

Previous studies have also considered the effect of individual differences, and especially gender, in betrayal aversion. Bohnet and Zeckhauser (2004) find evidence of substantial betrayal aversion in both genders, while Hong and Bohnet (2007) seem to find little evidence that women are betrayal averse. The three survey studies of forgiveness after betrayal in Finkel et al. (2002) come to three different conclusions: one finds men are more likely to forgive betrayal than women; another suggests less likely; and a third finds no difference. In a recent survey paper, Croson and Gneezy (2008) argue the balance of the evidence suggests women seem less willing to trust than men. This might suggest that women exhibit greater betrayal aversion than men.

Finally, the possibility of betrayal underlies much of the reputations literature (e.g., Houser and Wooders, 2006), and has also received attention within neuroeconomics. A notable example is Kosfeld et al. (2005), which draws a connection between oxytocin and trust decisions, arguing that oxytocin helps humans to overcome trust obstacles including betrayal.

Our research is novel in that it provides the first compelling evidence that investors prefer to avoid knowing whether a trustee chose to betray them, and when such knowledge is
unavoidable they are much more likely not to make trusting decisions. The next section details the experiment design we used to obtain these conclusions.

3 Experiment Design

Our goal is to create an experiment that rigorously identifies the effect of betrayal aversion on propensities to form trust relationships in economic exchange environments. To do this we consider three treatments of a one-shot binary investment game (or trust game, see Figure 1). In our game a human investor in room A (player 1) and a human trustee in room B (player 2) make decisions simultaneously. The decision alternatives presented to investors vary by treatment, as described below. Appendix A through appendix D provide transcripts of the instructions for all treatments.

3.1. KNOW Treatment

In the “KNOW” treatment each investor chooses either “trust” or “don’t trust”, and the trustee simultaneously chooses either “betray” or “reciprocate.” If the investor chooses “don’t trust” then both players earn five dollars regardless of the trustee’s decision. If the investor chooses “trust” then payoffs depend on how the trustee chooses to split $30. In particular, if the trustee chooses to reciprocate then each subject receives 15 dollars, while if the trustee chooses to betray then the investor receives two dollars and the trustee receives 28 dollars.

3.2. OPTION-TO-KNOW Treatment

The second treatment, the “OPTION-TO-KNOW” (OPTION) treatment, is the same as KNOW except each investor has the additional option to be paid according to a computer’s draw. The computer’s draw is not just a 50/50 chance of betrayal or reciprocate but is a draw from a pool identical to that specific session’s pool of trustees’ decisions. We describe below exactly how the computer draw occurs.

Note that this game tree was not distributed to subjects, nor were the terms “trust”, “betray”, or “reciprocate”, used to describe the game. The instructions also contain neutral framing.
Investors know that they are randomly assigned to their computer “decision” at the same time as they are randomly assigned to their human counterpart (see procedures below.) An investor in this treatment has three alternatives: (i) “don’t trust”, and thus assign five dollars to both one’s counterpart and oneself; (ii) “trust” and have one’s counterpart’s decision determine one’s payoff; (iii) “trust” and have a computer generated “betray” or “reciprocate” decision determine one’s payoff. Note that in the second and third cases the investor is aware that her trustee receives payment according to the trustee’s own decision. Consequently, a trustee’s earnings are independant of whether the investor chose payment based on the human counterpart’s decision or the computer’s decision, and are independent of the computer’s decision.

The third treatment, the “DONTKNOW” treatment, is the same as the KNOW and OPTION treatments except the investors choose between two alternatives: the “don’t trust” option (the $5/$5 split) and the computer option described above. Investors’ human trustees make the same decision as in the KNOW and OPTION treatments, and are paid in the same manner as well.

The instructions (truthfully) inform investors (see “Computer’s Decision” section in appendix A) that the computer makes its decision as follows. First, a computer tallies the number of “betray” and “reciprocate” decisions the N trustees made in that particular session. Next, the computer randomly assigns the numbers one through N either “betray” or “reciprocate.” Thus, there are exactly the same number of “betray” and “reciprocate” decisions in the N computer assignments as in the trustees’ data. It follows that the probability a trusting investor receives a “reciprocate” decision from her (randomly assigned) trustee is identical to the chance that the investor receives a (randomly assigned) “reciprocate” decision from the computer.

Note that each computer decision is not just the average human trustee decision. The computer decision can be thought of as a different draw, without replacement, from a pool of outcomes identical to the human trustee decisions, with the exception that there is no specific trustee associated with each outcome. As indicated in the investor’s instructions some computer

---

7 As the trustees’ instructions say that a possible scenario is for the trustee’s decision to determine both subjects’ payoffs, in DONTKNOW sessions one of the investor subjects was given a KNOW or OPTION set of instructions in a different room. For this reason, we were able to use identical trustee instructions in all treatments. The decisions from the few separated subjects have not been analyzed because their environment is not comparable to that of the other OPTION and KNOW treatments’ participants.
decisions are “betray” and some are “reciprocate”, in numbers exactly equal to the number of trustees who chose “betray” and number of trustees who chose “reciprocate” respectively.

3.3. Discussion of Design

It is worth emphasizing several features of our design. First, it may initially appear as if there is no functional difference between our treatment designs, that the computer decision is just the computer randomly assigning the investor to another anonymous person. The key is that as the computer is not assigning investors a new decision based upon any specific trustee, this draw carries with it no information about any individual trustee’s decision. Receiving a low payout from the computer informs the investor that some trustee chose to betray their counterpart. This information does not inform the trustee which trustee made that decision nor how many other trustees made that decision. Most importantly, an investor choosing to be paid by their computer decision does not learn what their own trustee counterpart’s decision is. The difference between our treatments lies with the information subjects learn from payment outcomes in each treatment, and as such, any treatment differences are not irrationality but revealed preferences over information content.

Second, as discussed above, we are interested in isolating betrayal aversion that subjects have to betrayal by an individual. We therefore designed the experiment to allow subjects only to avoid exposure to the knowledge that their own counterpart betrayed trust, but the design does not allow subjects to avoid exposure to the knowledge that someone’s counterpart trustee chose to betray trust. We therefore expect betrayal aversion to have an even larger effect on trust decisions than indicated by our data.

Third, the procedures used and the information and instructions given to trustees are identical among treatments (appendix B,) and trustees were in all cases drawn from the same population using identical recruiting procedures. Thus, we do not expect trustee decisions to differ systematically between treatments. Also, in an effort to ensure investors’ subjective expectations about trustees’ decisions displayed no systematic variation across treatments, we always gave trustees’ instructions to investors. Controlling subjective expectations has the twin advantages that (i) investors’ strategic uncertainty should not vary systematically among treatments; and (ii) subjects’ reference points are not likely to differ systematically across
treatments. Thus, loss aversion is not a plausible explanation for systematic differences in investor decisions among treatments.

It is also important to emphasize that trustees always earn what they choose if the investor chooses “trust”, regardless of whether the investor chooses the human or computer “trust” option. Thus, in relation to choosing not to trust, expected aggregate earnings increase when the investor chooses “trust.” However, the expected increase is invariant to which “trust” option the investor chooses. Because our design ensures subjective investor expectations are invariant among treatments, it follows that the only systematic difference among our conditions is whether one can avoid knowing that one’s specific trustee chose to betray.

3.4. Procedures

Upon arriving to the experiment subjects check in and proceed, as directed, to one of two rooms, room A (for investors) or room B (for trustees.) An equal number of subjects, say N, are seated in each room. (In our sessions, either eight or ten subjects were in each room.) Once seated, subjects read instructions and the experimenter reads the instructions aloud. Participants then answer a short quiz (see Appendix C and Appendix D) to ensure they understand the environment. After all participants successfully complete the quiz, each investor draws a number, without replacement, from a box containing the integers one through N. The outcome of that draw pairs the investor with a trustee previously randomly assigned that number. In conditions with a computer decision option, the same random draw pairs each investor to one of the N randomly generated computer decisions. After this, the experiment proceeds according to the treatment conditions. After privately receiving their results, subjects respond to a general questionnaire. Subjects receive payment in cash, privately, immediately prior to leaving the laboratory.

4 Hypotheses

Our main hypothesis, that people are betrayal averse, places three clear a priori restrictions on data from our treatments. These are as follows.
**Hypothesis 1:** *More investors will choose not to trust in the KNOW condition than the DONTKNOW condition.*

Economic theory not incorporating betrayal aversion predicts an equal fraction of investors fail to trust in both KNOW and DONTKNOW. The DONTKNOW treatment gives a baseline investor willingness to accept the pure gamble when betrayal aversion is not a consideration (an environment of state-uncertainty.) Our hypothesis that people are betrayal averse implies that adding the possibility of betrayal in the KNOW treatment (an environment of strategic-uncertainty) reduces investors’ willingness to trust in relation to DONTKNOW.

**Hypothesis 2:** *The frequency of trust decisions will be at least as high in the combined OPTION-TO-KNOW treatment as found in the higher of the KNOW or DONTKNOW treatments.*

Because Hypothesis 1 is that trust in DONTKNOW should be higher than found in KNOW, we can restate this hypothesis as the frequency of trust in the combined OPTION treatment is expected to be at least as great as found in DONTKNOW. The reason is that one can always choose to play the DONTKNOW game in the OPTION treatment, and similarly KNOW is available by simply ignoring the other alternative. Thus, the frequency of trust in OPTION should not be lower than is found in the greater of the other two treatments.

**Hypothesis 3:** *The frequency of trust in the KNOW condition will be at least as great as the frequency of “human” option trust in the OPTION-TO-KNOW condition.*

If people are on average betrayal averse then they should, on average, prefer to have the computer determine their return instead of a specific human. Note that betrayal aversion does not require all trusting investors to choose the “computer” option in the OPTION treatment. A reason is that some investors in the OPTION treatment might not exhibit betrayal aversion and thus can be indifferent between human and computer alternatives. People might also gain a positive utility from satisfying curiosity or from the knowledge that their counterpart reciprocated trust, both of which would lead to increased preference for the human trust option. These forces at least partially offset the expected disutility caused by betrayal knowledge to a betrayal averse investor, and will thus lead to underestimation of the percentage of betrayal

---

8 Based on where they sit in the lab.
averse subjects. Therefore, any increase in the percentage of investors observed trusting, between 
OPTION and KNOW, indicates only the lower bound for the percentage of betrayal averse 
investors in the sample.

5 Results

The experiments took place at the Interdisciplinary Center for Economic Science (ICES) at 
George Mason University. The randomly recruited subjects from the George Mason student 
body had no experience with trust games. In addition to any amount earned in the experiment, 
each subject received seven dollars for arriving to the laboratory on time. Subjects spent about 
40 minutes in the laboratory.

We report data from a total of 154 subjects in 77 investor-trustee pairs, with 26 pairs in 
each of KNOW and OPTION, and 25 pairs in DONTKNOW. Figure 2 describes the behavior 
of trustees by treatment. The fraction of trustees choosing to betray is 69.23%, 67.86%, and 
61.54% in the KNOW, DONTKNOW, and OPTION treatments respectively. Mann-Whitney 
tests reveal no significant difference in percentage of subjects choosing to betray. Overall, 
66.25% of trustees chose to betray, implying an empirical expected value to trust of $6.39 as 
compared to $5.00 for choosing not to trust.

Result 1: Significantly more investors chose to trust in the DONTKNOW treatment than 
trust in the KNOW treatment.

This result supports Hypothesis 1. Figure 3 describes the behavior of investors in the 
three treatments. Unlike trustees, significant differences in investor behavior between treatments 
are apparent. In KNOW, where trusting requires an investor to learn whether their counterpart 
chose to betray, we observed 65.38% of investors choosing trust. In DONTKNOW, where an 
investor cannot know whether her counterpart chose to betray, 92% of investors chose to trust. 
This difference is statistically significant (p<0.03) 10.

---

9 As noted in fn 5, the single “separated” investor in each DONTKNOW session (the one who participated in the 
KNOW or OPTION game) is excluded from our analysis. Our analysis does not exclude any trustee because all 
trustees in all treatments were in the same situation.

10 All of the p-values reported in this section are from two-sided Mann-Whitney tests.
Result 2: Investors trust significantly more in the OPTION-TO-KNOW treatment than in the KNOW treatment.

This result supports Hypothesis 2. In OPTION, where an investor can choose to avoid knowing whether she was betrayed, we found that 100% of investors chose to trust. This is significantly greater than the 65.38% who trusted in KNOW (p<0.01). Note that the fraction of investors who chose the trust gamble in DONTKNOW, 92%, does not statistically differ from the 100% who did so in OPTION (p=0.14).

Result 3: Significantly more investors chose to trust their human counterpart in the KNOW treatment than in the OPTION-TO-KNOW treatment (p <0.01).

This result supports Hypothesis 3. Compared to the 65.38% of investors who chose to trust their human counterpart in KNOW, in OPTION we observe that 46% of investors chose to trust the counterpart while 54% chose the computer trust option (see Figure 4). This result can be interpreted as suggesting that for about a fifth of people (65.38% - 46% = 19.38%) the expected benefit of the monetary gamble exceeds the expected disutility associated with betrayal aversion. Conversely, result 1 (which shows the difference in trust between DONTKNOW and KNOW) suggests that for about a fourth of people (92% -65.38% = 26.62%) the disutility associated with betrayal aversion exceeds the expected benefits of the monetary gamble. Overall, our three treatments suggest that at least 46% (26.62%+19.38%) of subjects hold sufficient betrayal aversion to influence economic decision making in our trust environment.

We found that the mean earnings among pairs of subjects to be $30 in OPTION and $28.06 in DONTKNOW, but only $23.64 in KNOW. The difference between KNOW and DONTKNOW is statistically significant (p<0.08), as is the difference between KNOW and

11 Note that if an investor trusts then the expected earnings of a pair of subjects is always 30 dollars regardless of the treatment. If an investor chooses a computer trust option the realized earnings of the pair could be either 17, 30, or 43 dollars, while the expected earnings from trust remains 30 dollars. We use expected earnings from trusting in a particular session for the earnings of an investor in order to have accurate reporting.
OPTION (p<0.02). Thus, economic efficiency falls when the possibility of learning of a betrayal is unavoidable.

**Result 4:** *In the KNOW treatment, pairs with female investors earn significantly less than pairs with male investors.*

Of the 77 investors who participated in our experiment, 45 were male and 32 female. In KNOW, mean earnings among pairs with a male investor is $27.55 (n=11), an amount significantly greater than the $20.78 average (n=15) among pairs with female investors (p=0.05). That we obtain statistical significance with our rather small sample may suggest that women exhibit greater betrayal aversion than men in this game. We find no significant gender effects on earnings within the OPTION or DONTKNOW treatments. Also, we find no evidence of gender effects in trustees’ decisions to betray.

**6 Discussion**

We present, to our knowledge, the first rigorous evidence that failures to trust can be traced to betrayal aversion as distinct from risk-aversion or other factors that appear in standard economic theory based on selfish expected earnings maximization. Our results indicate that less than half of investors choose to know their trustee’s decisions when they can avoid such knowledge. Moreover, when such knowledge is unavoidable, a significant fraction simply opt-out by choosing not to invest. Thus, betrayal aversion has a statistically and economically significant detrimental impact on economic efficiency.

Some have noted that our data could be explained if, somehow, the presence of a computer and the associated instructions (as in our DONTKNOW and OPTION treatments) increases one’s willingness to invest in relation to cases when a computer is not present (as in our KNOW treatment). To explore this possibility, we ran twenty subject pairs through a treatment similar to KNOW but with a computer present (the KNOW2 treatment.) As in the

---

12 The questionnaire responses from investors to the question “How would you feel if your counterpart chose D?”, seem to support our view that subjects’ attitudes are consistent with betrayal aversion. Investors reported they would feel “angry”, “miffed”, “annoyed”, “sad” or “betrayed” if their trust was not reciprocated. One subject, who chose the computer option in OPTION, said betrayal by a human would leave him feeling, “[o]ffended, thus I didn’t choose that option.” On the other hand, a subject who “trusted” in DONTKNOW replied that if he did not receive
OPTION and DONTKNOW treatments, the KNOW2 treatment matched investors with a computer decision and required the investors to understand how the computer made its decisions. However, the computer’s decisions did not affect investors’ decisions. We found 17 of 20 subjects to trust, a frequency insignificantly different from that in the original KNOW treatment (p = 0.14), but statistically significantly lower than trusted in the OPTION treatment (p=0.04). Our evidence that the presence of a computer does not in itself lead to significant increases in trust supports the importance of betrayal aversion in explaining our data.

Our experimental environment bears some similarity to online anonymous exchange environments (such as eBay, Amazon, etc.) where the identity of trading partners is either anonymous or limited to email addresses or a “store” name. Just as people are averse to betrayal in our experiments, it seems reasonable to expect that people would be averse to betrayal in these online markets as well. Past studies on reputation effects (e.g., Livingston, 2005; Houser and Wooders, 2006) demonstrate that sellers with better reputations earn higher prices. While previous studies have explained this result within the context of state-uncertainty, our findings point to the potential importance of strategic-uncertainty in this environment, and suggest an alternative explanation for reputation premiums: their value lies not only in reducing state-uncertainty but also in reducing the chance one experiences a negative emotion.

Our data indicate people prefer to trade in environments where it is difficult to know that a trusted counterpart chose to betray. This preference might provide a partial explanation for the emergence of posted-offer markets. For example, Saturn made its name by offering “No Hassle” and “No Haggle” sales, ensuring that all customers received the same products at the same price. This policy largely eliminates a consumer’s strategic uncertainty associated with a negotiation, allowing them to avoid knowing they were “taken” or “swindled” by their car salesperson. The policy does nothing, of course, to reduce any state-uncertainty associated with the quality of the car. More generally, the evolution from small, neighborhood oriented “mom and pop” stores to large regional supermarkets shift the environment away from strategic-uncertainty and in the direction of state-uncertainty. In particular, one does not expect the cashier at a box-store who processes transactions to be aware of the quality of the purchased products.

the higher payoff, “I would feel neutral because it really is the computer which decides what letter I’m assigned,” and another indicated they would feel “Nothing, as my earnings are decided by computer.”
We noted that a natural explanation for an investor’s preference not to know her specific trustee’s decision is a preference to avoid the possibility of experiencing a negative emotional outcome. In the psychology literature this is referred to as “situational modification,” and is part of so-called “antecedent-focused emotion regulation” (Gross 1998). A more complete understanding of how expectations mediate economic decision-making will emerge from an investigation of this area. Additionally, it remains an open question whether the negative emotional response to betrayal stems from the knowledge that one’s specific counterpart betrayed trust or stems from the knowledge that one is the victim of betrayal from one’s counterpart. Emotion regulation studies may find studying the differences between these two types of betrayal aversion profitable.

Evolution has endowed people with a natural tendency to approach social exchange situations with caution. Future research aimed at understanding the evolutionary benefits of betrayal aversion would be profitable. In light of the results of Kosfeld et al. (2005) we speculate that Oxytocin might have co-evolved with betrayal aversion in order to amplify humans’ willingness to trust in intimate social relationships. Such co-evolution would help to explain why trust can emerge in contexts where betrayal would be especially painful (e.g., trust placed in a dear friend), and yet can fail to develop in less intimate economic-exchange environments. We look forward to exploring this and related hypotheses in future research.

Acknowledgements: We thank Ernst Fehr and Benjamin Schoefer for a valuable discussion that led to the experiment design reported in this paper. For helpful comments and suggestions, we thank Omar Al-Ubaydli, Vernon Smith, Virgil Storr, Bart Wilson, Erte Xiao, Richard Zeckhauser and participants in research seminars at ICES, the Mercatus Center, George Mason University, University of Lyon, University of Munich, University of Zurich, the International ESA 2008, and the SEA meetings 2008. Kail Padgitt and Adam Smith provided capable laboratory assistance. We gratefully acknowledge research support from the International Foundation for Research in Experimental Economics and the National Science Foundation Research Grant SES-0851250.
References


Figure 1
Investment Game

Player 1
(Investor)

Don’t Trust
$5
$5

Trust

Betray

Player 2
(Trustee)

Reciprocate

$2
$28

$15
$15
Figure 2
Betrayal By Treatment

Trustees Who Betrayed

- **KNOW**
  - N = 26
  - 69.23%

- **DONT-KNOW**
  - N = 28
  - 67.86%

- **OPTION-TO-KNOW**
  - N = 26
  - 61.54%

Treatment

- ■ Trustees Who Betrayed
Figure 3
Trust By Treatment

- KNOW: N = 26 (65.38%)
- DONT-KNOW: N = 25 (92.00%)
- OPTION-TO-KNOW: N = 26 (100.00%)

Treatment
- ■ Investors Who Trusted
Figure 4
Type of Trust Decision
By Treatment

- Trust With Betrayal Knowledge
  - KNOW
    - N = 26
    - 65.38%
- Trust Without Betrayal Knowledge
  - OPTION-TO-KNOW
    - N = 26
    - 46.15%

- Trust With Betrayal Knowledge
- Trust Without Betrayal Knowledge
Appendix A

Room A (Investors) Instructions: OPTION-TO-KNOW Treatment

Thank you for participating in today’s experiment. You’ve earned a $7 show-up bonus for participating. In reading and following the instructions below, you have the potential to earn significantly more. You have been randomly assigned to Room A. You will also be randomly and anonymously assigned to a person in Room B. Your counterpart will not be told your name, and you will not be told his/her name.

How you are matched with your counterpart:
Each of the 10 Room A persons will be matched with a different Room B counterpart for the entire experiment. The experimenter will bring around a box with the numbers 1 through 10 inside. The number you draw will assign you to one of the 10 counterparts in Room B (B1 through B10 coinciding with the numbers 1 through 10 in the box). The number also matches you with one of the 10 computer number decisions (coinciding with numbers 1 through 10 in the box).

Your Decision:
You have three options for how the earnings for you and your counterpart will be determined in today’s experiment. You must choose exactly one of the following three options:

- You receive $5 and your counterpart receives $5.
- Both you and your counterpart are paid based on his/her decision between “U” ($15 for you and $15 for him/her) and “D” ($2 for you and $28 for him/her).
- Your counterpart is paid according to his/her decision between “U” and “D”, and you are paid based on a computer’s choice between either “U” or “D”.

You will not be told what the computer’s decision was, or what your counterpart’s decision was, unless you choose that earnings option.

---

\[13\] DONTKNOW treatment did not include the second payment choice option. The KNOW treatment did not include third payment choice option and did not include the “computer’s decision” paragraph (next page.) The KNOW2 treatment did include the “computer’s decision” paragraph.
**Room B Decision:** (The instructions given to your counterpart)
You will be anonymously assigned to a Room A counterpart who drew your number randomly from a box with the numbers 1 through 10 inside. This person will be your counterpart for the entire experiment. Your counterpart will make a decision that can affect your earnings in today’s experiment. He or she can choose for both of you to be paid $5. Another possibility is that he/she will let you determine both of your payoffs. If he/she chooses this option and you choose “U”, then you get paid $15 and he/she gets paid $15. If you choose “D”, then you get paid $28 and he/she gets paid $2. Your payoff will be determined in one of these two ways. Your counterpart can choose only one of the earnings methods. We will ask you to make your decision on “U” or “D” at the same time that your counterpart makes his or her choice. Your decision will only determine your payoff if your counterpart did not choose the option to give you $5.

**Computer’s Decision:**
After the Room B participants make their decisions, the computer will assign either "U" or "D" to each of the ten numbers. The computer has been programmed to assign dollar values to each of the 10 numbers in the box according to the decisions made by the Room B participants. What this means is that the number of "U" choices made by the computer is exactly the same as the number of "U" choices made by the participants in room B. Also, the number of "D" choices made by the computer is exactly the same as the number of "D" choices made by the room B participants. (Note: while the number of “U” numbers and number of “D” numbers are the same as in the Room B decisions, which numbers are assigned “U” or “D” is randomly decided by the computer) For example: if five Room B participants choose "U", then five of the numbers between 1 and 10 are randomly assigned to have the "U" payoff, and the remaining five numbers are assigned to the "D" payoff. (Note: the numbers used here are only an example and not necessarily representative of Room B decisions)
Appendix B

Room B (Trustees) Instructions: Every Treatment

Thank you for participating in today’s experiment. You’ve earned a $7 show-up bonus for participating. In reading and following the instructions below, you have the potential to earn significantly more. You have been randomly assigned to Room B. You will also be randomly and anonymously assigned to a person in Room A. Your counterpart will not be told your name, and you will not be told his/her name.

You will be anonymously assigned to a Room A counterpart who drew your number randomly from a box with the numbers 1 through 10 inside. This person will be your counterpart for the entire experiment. Your counterpart will make a decision that can affect your earnings in today’s experiment. He or she can choose for both of you to be paid $5. Another possibility is that he/she will let you determine both of your payoffs. If he/she chooses this option and you choose “U”, then you get paid $15 and he/she gets paid $15. If you choose “D”, then you get paid $28 and he/she gets paid $2. Your payoff will be determined in one of these two ways. Your counterpart can choose only one of the earnings methods. We will ask you to make your decision on “U” or “D” at the same time that your counterpart makes his or her choice. Your decision will only determine your payoff if your counterpart did not choose the option to give you $5.
Appendix C

Room A (Investors) Quiz Questions: OPTION-TO-KNOW Treatment

1) You can choose multiple methods to determine your earnings. True / False

2) Your counterpart’s earnings are the same whether you chose to be paid based on your counterpart’s earnings or you chose to be paid based on the computer decision. True / False

3) Your counterpart only gets earnings based on his/her decision if you do not chose for you both to be paid $5. True / False

4) If you choose to be paid based on your counterpart’s decision and he/she chose U, what are your earnings? _______ your counterpart’s? _______

5) If you choose to be paid based on your counterpart’s decision and he/she chose D, what are your earnings? _______ your counterpart’s? _______

6) How many numbers will the computer randomly assign the U value to if exactly 6 Room B counterparts choose U. _______

7) How many numbers will the computer randomly assign the D value to if exactly 6 Room B counterparts choose D. _______

8) The value of the computer decision assigned to a number may not be the same as the value of the decision made by the Room B counterpart with that number. True / False

9) Will you know what your counterpart chose if you choose earnings based on the computer decision? _______

10) Will you know what the computer decision was if you choose earnings based on your counterpart’s decision? _______
Appendix D

Room B (Trustees) Quiz Questions: All Treatments

1) If your counterpart chooses for you to both receive $5 and you chose D how much are you paid?________ your counterpart? _________

2) If your counterpart chooses for you to both receive $5 and you chose U how much are you paid?________ your counterpart? _________

3) If your counterpart chooses to be paid based on your decision and you chose D how much are you paid?________ your counterpart? _________

4) If your counterpart chooses to be paid based on your decision and you chose U how much are you paid?________ your counterpart? _________