09.01.2017

Plausible Deniability and Cooperation in Trust Games*

Anthony S. Gillies
Rutgers University

Mary L. Rigdon
Rutgers University

Abstract

What motivates agents to choose pro-social but dominated actions in principal-agent interactions like the trust game? We investigate this by exploring the role higher-order beliefs about payoffs play in an incentivized laboratory experiment. We find that when there are asymmetries in such higher order information that generates "plausible deniability", agents exploit that: otherwise trustworthy types are tempted into defecting.

KEY WORDS: trust, reciprocity, social preferences, trust game, guilt aversion, behavioral economics, higher order beliefs

JEL CODES: C91, D91, D82

1 Introduction

A(lex) and B(illy) each have \$10 for pizza. Sadly, the by-the-slice offerings for \$10 aren't great. But the whole pies (which neither can afford alone) are magnificent — in fact, they'd each enjoy it more than \$10-worth of utility. There is, however, a catch: while B is at the front of the line, A is at the back (and unlikely to get to the front before the pizzas are no more). It seems they are stuck. But if A could trust B, giving over her \$10 to B, then they would both

* Corresponding author: Rigdon (mrigdon@rutgers.edu). Authors' names are listed alphabetically. Thanks to Cary Deck, Dan Houser, Bart Wilson, and audiences at the Economic Science Association summer meetings, and the Nordic Conference on Experimental Economics for comments, and to Ning Nan, Yaye Aida Ba, and Patrick Vandenplas for research assistance in running the experimental sessions. This research was supported by a grant from the International Foundation for Research in Experimental Economics.

be much better off. Assuming, that is, that B doesn't just take the whole pie for herself.

The structure of the problem facing A and B is common and familiar: it is a principal–agent problem in which, since there is no third-party that can force B's hand to reciprocate A's trust, the agent's action is not contractable. That fact makes A's trust risky: if people are rational and self-interested then agents will pursue their own interest and principals, knowing this, won't enlist their help. Thus conventional game theoretic analysis favors no exchange at all (and sub-par pizza). This is inefficient since A and B would both prefer to share a pie. And it is at odds with the fact that substantial numbers of principals and agents (both in the laboratory and in the wild) do manage to reach cooperative outcomes that Pareto dominate the subgame perfect equilibrium, even in anonymous one-shot encounters. That makes understanding the motivations that drive this choice behavior both theoretically and empirically significant. We focus, in particular, on the agents in the exchanges. But we replace the question of why B cooperates with this one: What can tempt an otherwise trustworthy B to defect?

We are interested in what role information (both first- and higher-order) about the payoff structure of interactions plays in motivating pro-social choice behavior. Of course, in the example above, it is common knowledge what is at stake for A and B. But that does not always happen. Sometimes it is common information that A and B only have information about their own interests. In that case, B has no way to choose the pro-social option (except by accident) and knows that A knows this. And sometimes the information is asymmetric. Suppose B knows what's at stake for each of them, but knows that A only has information about what is at stake for her. In that case, B has plausible deniability: she knows that as far as A is concerned, B might have no way to non-accidentally choose the pro-social option. We report evidence from an incentivized laboratory experiment that suggests that for a significant portion of the population, the motivation to cooperate is related to whether or not they have plausible deniability: the presence of plausible deniability is enough to tempt otherwise cooperative types to eat the whole pie.

The rest of the paper is organized as follows. Section 2 contains an informal discussion of how the kinds of distributions of information we are interested in lead to different predictions of choice behavior in the environment we focus on. Section 3 formalizes this, and shows in a more precise way how to derive the hypotheses we test. The hypotheses are summarized in Section 4. Section 5

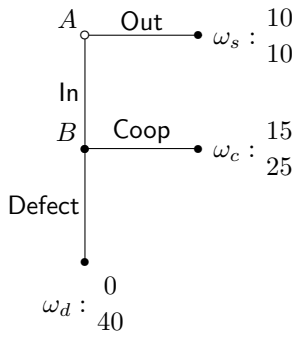


Figure 1: Trust Game

discusses our experimental procedures, and Section 6 contains our main results. We conclude in Section 7.

2 Plausible deniability and higher-order information

We focus on a simple two-person one-shot trust game (Figure 1).[1] The game has enough of the structural properties to make it a reasonable model of a principal–agent interaction under an implicit contract: A can enlist B, but B's choice at her decision node is beyond A's control and there is no enforcement agency or shadow of the future or punishment mechanism that can be relied on. But the game is also simple enough to allow us to easily manipulate subjects' higher-order beliefs about the space of monetary distributions in a play of the game. As is well attested: there is substantial off-equilibrium but efficient cooperation in such situations (see, for instance, Berg, Dickhaut & McCabe 1995; Ortmann, Fitzgerald & Boeing 2000; McCabe & Smith 2000; Bohnet & Zeckhauser 2004; McCabe, Rigdon & Smith 2002 and Camerer 2003; Johnson & Mislin 2011 for an overview).

Our main hypothesis is that plausible deniability can be a barrier to cooperation: B's willingness to cooperate is affected by her higher-order information and in particular by her information about A's information about B's information about the monetary payoffs in the subgame that B controls. Put the other way around: otherwise cooperative types are tempted to defect in the presence of plausible deniability. Notice that this means that we will need to

[1] We have decorated the terminal nodes both with labels and with payoffs for reasons that will be clear in Section 3.

model in a sensible way uncertainty about payoff information. We will return to this below.

We focus on three distributions of payoff information. First, as is completely standard, suppose payoff information is common knowledge between A and B. That is, suppose:

- A and B both know the full distribution of possible payoffs to each player;
- A and B both know that A and B know this;
- and so on ....

In that case, B may have reasons for choosing the pro-social Coop instead of the dominant action Defect: perhaps B is (to some extent) inequity averse, or B feels bound by an implicit contract between A and B, or B is guilt averse, or B has some other motivation. Just what those reasons are or might be is, of course, part of the question we are addressing here and why we are interested in other ways of distributing payoff information.

Second, consider the other extreme in which payoff information is completely private. That is suppose:

- A only knows what payoffs A faces at the terminal nodes (and has no other relevant information);
- B only knows what payoffs B faces at the terminal nodes (and has no other relevant information);
- this is common information between them.

Of course, in that case there is no reason for B to choose Coop.

The third (more complicated) case is a situation in which payoff information is asymmetric. Suppose:

- B knows the full space of payoffs for both A and B;
- A only knows A's payoffs (and has no other relevant information);
- B knows that this is the way information is distributed.

A has little to go on here, and B knows this. In fact, A should have no non-trivial beliefs about B's payoffs and no non-trivial beliefs about whether B has any information about A's payoffs. And, of course, B knows this. So after

B chooses, A can't conclude much of anything about much of anything — not about B's payoff, their comparative payoffs, or B's motives. In such a situation, B has plausible deniability: she knows that for all A knows, B might not know how B's decision impacts A. Is an otherwise trustworthy B tempted here to Defect rather than Coop?

Before we move on to showing how to model these distributions of information in a precise way that interfaces with existing theories of B's motivation, and how those distributions lead to concrete predictions for relative rates of cooperation and defection, we want to highlight three ways in which the results we report here are novel.

First, while there is a body of results pointing away from explaining off-equilibrium but pro-social behavior simply in terms of preferences for fairness or inequity aversion and toward some sort of higher-order belief-dependent or intensions-based explanation of cooperation in these or similar contexts (starting with McCabe, Rigdon & Smith 2003; Falk, Fehr & Fischbacher 2008), these tests all involve altering the strategic landscape of the standard trust game (principally by altering the choices for or payoffs to A).[2] Our design does no such thing: we manipulate only the way in which information about the payoffs is distributed between A and B.

Second, while there are experimental probes of guilt aversion (Charness & Dufwenberg, 2006; Battigalli & Dufwenberg, 2006), our target here is a broader class of explanations of which guilt aversion is a special case. We are interested specifically in exploring whether higher order beliefs can motivate B to cooperate and whether an asymmetry in those higher order beliefs can crowd out cooperation. While that is in principle compatible with guilt aversion explanations of strong reciprocity, it is also compatible with explanations that trade on implicit contracts and inter-personal mental accounting (McCabe & Smith, 2001; McCabe et al., 2003; Smith, 2008; Smith & Wilson, 2014).

Third, previous experimental probes of guilt aversion have focused on non-binding communication between players as a means of manipulating the relevant degrees of belief (Charness & Dufwenberg, 2006). Our experiment does no such thing; we can thus finely control who has the higher-order information and who

[2] A few examples: McCabe et al. (2003) trim the trust game by removing A's outside option, Snijders & Keren (2001) vary payoffs creating smaller or larger gaps in A's opportunity cost, and Smith & Wilson 2017 introduce downstream punishment options for B (if A doesn't trust) or A (if B doesn't reciprocate). Dana, Weber & Kuang (2006) probe plausible deniability in dictator giving. However, they do so by in effect introducing uncertainty about where in the game tree A finds herself (by giving Nature a chance to divide the allocation).

does not.

3 Modeling higher-order information

Since our interest is in B's behavior, let's focus on the subgame of the standard trust game that B controls. We want to more precisely characterize the kind of higher-order information that may motivate B, give a simple characterization of plausible deniability, and draw out some hypotheses.

Our goal here is to show how, given an existing model of B's motivation, we can derive predictions from that model for B's choice across information conditions. However, a key ingredient for us is that agents may be uncertain (and have higher order beliefs about that uncertainty) about the distribution of payoffs the agents face at the terminal nodes. This uncertainty is not something easily represented in existing models.[3] So we develop a theory neutral framework for representing it (using simple Aumann structures), and then re-cast existing models in it to derive hypotheses.

We assume that players have perfect information about the space of actions open to each of them, but we don't assume that that necessarily holds for information about monetary payoffs. To represent uncertainty about payoff information, first, decorate each terminal node with a label ω. Assume that, for any terminal node, A's (B's) possible monetary payoffs must lie between 0 and Π_A (Π_B). A state s is a function from terminal nodes to monetary payoffs: $s(\omega) \subseteq \Pi_A \times \Pi_B$. We use $s, t, \ldots$ (with or without subscripts) as variables over states. For any state s and terminal node ω, let $\pi_i(s, \omega)$ be the monetary payoff in s that i faces at ω.[4]

Subsets of S are events.[5] So, for example, the event that A stands to earn 10 at the (outside option) node ω_s is $\{t\colon\ \pi_A(t, \omega_s) = 10\}$. An event e obtains at a state s iff $s \in e$.

[3] Attanasi, Battigalli & Manzoni 2016 develop a model of guilt aversion in incomplete (psychological) games, but there too it is assumed that information about material payoffs is part of the structure of the game and therefore common knowledge. Our interest lies precisely in separating terminal nodes (equivalently: complete paths through a game tree) from the material payoffs at terminal nodes: there can be perfect information about the first without there being perfect information about the second.

[4] That is: $\pi_A(s, \omega) = s^1(\omega)$ and $\pi_B(s, \omega) = s^2(\omega)$ where $s^i(\omega)$ is the ith projection of $s(\omega)$.

[5] We have simplified the presentation by only having events be events about what the possible payoffs are. So they do not represent information about which distribution will be reached. That information isn't necessary given our purposes.

An event that we will return to frequently is the event that characterizes A's payoff possibilities in the standard trust game (Figure 1). Let P_A be the event that A stands to earn 10 at ω_s, 15 at the cooperative ω_c, and 0 at ω_d:

$$P_A = \{t\colon\ \pi_A(t, \omega_s) = 10 \text{ and } \pi_A(t, \omega_c) = 15 \text{ and } \pi_A(t, \omega_d) = 0\}$$

Similarly let P_B be the event that B stands to earn 10 at ω_s, 25 at the cooperative ω_c, and 40 at ω_d:

$$P_B = \{t\colon\ \pi_B(t, \omega_s) = 10 \text{ and } \pi_B(t, \omega_c) = 25 \text{ and } \pi_B(t, \omega_d) = 40\}$$

Finally, let's distinguish s^* as the actual state:

$$P_A \cap P_B = \{s^*\}$$

That is, s^* is the unique state in which the payoffs each player faces are as in the standard trust game above.

If we only wanted to think about first-order information about monetary payoffs, then we could stop here. But we also want to represent events about the distribution of high-order information about monetary payoffs — for instance, the event that A doesn't know that B knows what A stands to earn at the (cooperative) node ω_c. So we represent players' knowledge using structures $\langle S, I\rangle$ where S is the set of states and I is a function from players to partitions S_i of S. We write $S_i(s)$ to denote the cell of the partition S_i that has state s as a member. Intuitively, i's knowledge in state s is just what is true throughout $S_i(s)$. Notice that if $t \in S_i(s)$ then t is compatible with what i knows in s. Slightly more formally: where e is any event, the event that i knows that e is represented this way:

$$\mathsf{K_i}(e) = \{s\colon\ S_i(s) \subseteq e\} \tag{1}$$

So i knows that e in a state s iff $s \in \mathsf{K_i}(e)$.

We will also assume that B has coherent degrees of belief over states compatible with what she knows: $p_B(s)$ represents her credence that $s^* = s$ and, for any event e, $p_B(e) = \sum_{s \in e} p_B(e)$.

Return now to the event P_A characterizing A's (actual) payoff possibilities. Since it is an event, it is the sort of thing that can be known by an agent. For instance: given a state s, B knows P_A iff $S_B(s) \subseteq P_A$. This, too, is an event and so it too can be known — the result is some event characterizing higher-order knowledge about payoff information. For instance: B knows at s that A knows that B knows P_A iff $s \in \mathsf{K_B K_A K_B} P_A$.

Common knowledge of an event is defined in the expected way: the event e is common knowledge between between A and B iff both know that e, both know that they know it, both know that both know that they both know it, and so on. Let $\mathsf{E}(e)$ be the event that both A and B know that e. Then common knowledge is the intersection of these events, all the way up the hierarchy of "both know that":

$$\mathsf{C}(e) = \bigcap_{k=1}^{\infty} \mathsf{E}^i(e) \tag{2}$$

where $\mathsf{E}^1(e) = \mathsf{E}(e)$ and $\mathsf{E}^{n+1}(e) = \mathsf{E}^1\mathsf{E}^n(e)$.

We are interested in three ways of distributing information about the events P_A and P_B — the events characterizing the payoff possibilities in the standard trust game — between A and B. When it comes to our experimental design, these information conditions will be our treatments:

COMMON P_A and P_B are both common information between A and B:

$$\mathsf{C}(P_A \cap P_B)$$

PRIVATE A only knows P_A and B only knows P_B:

$$\mathsf{C}((\mathsf{K_A}(P_A) \cap \overline{\mathsf{K_A}(P_B)}) \cap (\mathsf{K_B}(P_B) \cap \overline{\mathsf{K_B}(P_A)}))$$

ASYMMETRIC A only knows P_A and B knows both P_A and P_B and knows that A doesn't know that B knows both P_A and P_B:

$$\mathsf{K_A}(P_A) \cap \mathsf{K_B}(P_A \cap P_B \cap \overline{\mathsf{K_A}(\mathsf{K_B}(P_A \cap P_B))})$$

Note that in PRIVATE if B chooses Defect she is blameless: she has no information about how this might affect A. For all B knows, choosing Defect over Coop might bring a windfall to A. In ASYMMETRIC, B has plausible deniability: although she knows how her choice impacts A, she knows that A doesn't know this. Why might plausible deniability make a difference to whether B chooses Defect or Coop? To get a better idea how this might make a difference, we quickly survey three sorts of theories of what might motivate B and show how when put into the current framework they treat higher-order payoff information in the trust game, thereby deriving our hypotheses.

3.1 Homo economicus

As an extreme example, consider the classical maximizing theory of B: B simply chooses whichever action maximizes her expected utility, where B's utility is solely a function of B's monetary payoffs. B's subjective expected utility (SEU) of choosing an action is her expectation of the payoff associated with that action. So where p_B is B's distribution of degrees of belief over the states compatible with what she knows in s^*:

$$SEU(\mathsf{Defect}) = \sum_{s \in S_B(s^*)} p_B(s)\pi_B(s, \omega_d) \tag{3}$$

$$SEU(\mathsf{Coop}) = \sum_{s \in S_B(s^*)} p_B(s)\pi_B(s, \omega_c) \tag{4}$$

Now, if B knows the payoffs B faces — that is, $S_B(s^*)$ contains states only if they differ at most in what monetary payoff they say A faces — and B's degrees of beliefs are uniform over $S_B(s^*)$ (since she has no other relevant information) then a maximizing B will choose Defect. It is easy to check that for every distribution of payoff information we are interested in (COMMON, ASYMMETRIC, PRIVATE) this minimal condition on what B knows is met: in each case, $S_B(s^*) \subseteq P_B$.

That gives us a concrete hypothesis: a *homo economicus* B will choose Defect regardless of how payoff information is distributed. Put another way: if the population is made exclusively of *homo economicus* player types, then rates of defections should be high and constant no matter how higher-order information about monetary payoffs is distributed between players.

3.2 Homo altruans

That is an extreme example. A more plausible hypothesis is that B may be motivated by a taste for fairness or a distaste for inequity. Such a B chooses so as to maximize some adjusted expected utility, where the adjustment reflects how much she disprefers unequal or unfair outcomes. Theories like this trade on allowing B's expected utility to be a function of more than simply what B believes about B's monetary payoffs. It is also a function of what B believes about A's monetary payoffs. But — and this is our point here — such theories do not naturally make utilities sensitive to higher-order beliefs about monetary payoffs and so do not naturally have room to say that some agents are motivated

by such beliefs. To illustrate the basic point, we again focus on a simple example: (a simple extension of) the Fehr & Schmidt 1999 inequity aversion model. B must choose between Defect and Coop. The hypothesis is that B's utility at a terminal node depends on what payoff B faces at that node and what payoff A faces at that node. So B's expected utility of reaching a node is that adjusted utility weighted by B's degree of belief that that node will be reached.

This model adjusts utilities simply by subtracting in proportion to inequity. There are two ways for outcomes to be unequal, so there are two ways to weight the utilities. Let's roll them into one inequity measure: $\Delta_B(s, \omega)$. This is how much B's monetary payoff is affected by an unequal or unfair final division at terminal node ω in state s:

$$\Delta(s,\omega) = \begin{cases} \alpha(\pi_A(s,\omega) - \pi_B(s,\omega)) & \text{if } \pi_A(s,\omega) \geq \pi_B(s,\omega) \\ \beta(\pi_B(s,\omega) - \pi_A(s,\omega)) & \text{otherwise} \end{cases} \tag{5}$$

where, as in the basic Fehr & Schmidt model, $\alpha \geq \beta$ and $0 \leq \beta \leq 1$.

Now it is straightforward to give B's inequity averse or fairness expected utility (FEU) in terms of this measure. B's FEU of choosing Coop in the actual state s^* is the weighted average of B's fairness-adjusted utilities across the states compatible with what she knows in s^* where the weights are her degrees of beliefs:

$$FEU(\mathsf{Coop}) = \sum_{s \in S_B(s^*)} p_B(s)(\pi_B(s,\omega_c) - \Delta(s,\omega_c)) \tag{6}$$

where p_B is B's probability distribution over the states in $S_B(s^*)$. Similarly for Defect:

$$FEU(\mathsf{Defect}) = \sum_{s \in S_B(s^*)} p_B(s)(\pi_B(s,\omega_d) - \Delta(s,\omega_d)) \tag{7}$$

Note that if $s^* \in \mathsf{K_B}(P_B) \cap \overline{\mathsf{K_B}(P_A)}$ — that is, if B only knows what her payoff opportunities are — then for every $s, t \in S_B(s^*)$, $\pi_B(s,\omega_c) = \pi_B(t,\omega_c)$ but in general $\pi_A(s,\omega_c) \neq \pi_A(t,\omega_c)$ and similarly for ω_d. If p_B is uniform over $S_B(s^*)$, maximizing these expected weighted utilities reduces to maximizing B's utilities simplicter: FEU reduces to SEU. Thus, given the payoffs in the standard trust game, if payoff information is distributed as in PRIVATE then $FEU(\mathsf{Defect}) \geq FEU(\mathsf{Coop})$ iff $40 \geq 25$. That is the right result: an inequity averse B should choose Defect if she only knows her own payoff opportunities and has no information about A's.

Of course, if B knows the relevant payoff information about A then things are different. For instance, in an information environment like COMMON, since $s^* \in \mathsf{C}(P_A \cap P_B)$ it follows that the only states compatible with what B knows (the only states in $S_B(s^*)$) are states in which the payoffs opportunities each player faces are the same as in the actual state s^*. Thus (6) and (7) reduce to the more familiar:

$$\begin{aligned} FEU(\mathsf{Coop}) &= \pi_B(s, \omega_c) - \Delta(s, \omega_c) \\ &= 25 - \beta(25 - 15) \end{aligned}$$

$$\begin{aligned} FEU(\mathsf{Defect}) &= \pi_B(s, \omega_d) - \Delta(s, \omega_d) \\ &= 40 - \beta(40 - 0) \end{aligned}$$

That is: the basic Fehr & Schmidt model comes out as a special case whenever B knows the payoff opportunities for both A and B. This is unsurprising for information environments like COMMON. But it is equally true for payoff environments like ASYMMETRIC: for here too it is easy to verify that $s^* \in \mathsf{K_B}(P_A \cap P_B)$. Although there are plenty of informational differences between COMMON and ASYMMETRIC, these differences are not visible to fairness theories of motivation like this.

That gives us a concrete hypothesis: a fairness-motivated B may cooperate in an environment like COMMON (depending on just what her β value is), but whatever she chooses we would expect that she should choose in exactly the same way in an environment like ASYMMETRIC. Put another way: if the population is made up only of homo economicus and homo altruans, then the rate of defection in COMMON and ASYMMETRIC should be equal and lower than the rate of defection in PRIVATE.

3.3 Homo reciprocans

So far none of the theories make room for plausible deniability making a difference to B's behavior. But suppose B's utility is a function, in part, of what she believes about what A believes about B's behavior. Perhaps B doesn't want to be seen as shirking on the implicit contract between A and B by not living up to its terms (Hoffman, McCabe & Smith, 1998; McCabe & Smith, 2001; Smith, 2004) or perhaps B has some form of guilt aversion and only feels guilty when she knows her actions let A down (Charness & Dufwenberg, 2006; Battigalli & Dufwenberg, 2006). In any case, these sorts of theories about why

B may cooperate make explicit appeal to B believing something about what A expects of B and so make plausible deniability relevant to what B chooses. We now show how to make this precise in our current framework and show how that implies a hypothesis for the distributions of information we are interested in. Again, we pick a simple implementation to make our point.

We start with a simple idea: B's motivation to go for Coop instead of Defect is a matter of whether her utility for Defect gets dragged down sufficiently by her feeling guilty.[6] There are two ingredients to how we will model this. First, as in standard models of guilt aversion, B's guilt will be a function, in part, of what B thinks A stands to lose if B opts for Defect instead of Coop. Second, we will weight B's guilt by her belief about the presence or absence of plausible deniability. It is easiest to model things based on B's belief about the absence of plausible deniability so that is what we do in what follows. We now define these ingredients and then show how to derive hypotheses about choice behavior given different distributions of payoff information.

Given a state s, B's belief about what A stands to lose if B chooses Defect instead of Coop is simply the difference in her expectations across states compatible with what she knows in s of A's payoffs in ω_c and ω_d. Formally, this is recorded in the variable $\delta_B(s,\omega)$:

$$\delta_B(s,\omega) = \sum_{t\in S_B(s)} p_B(t)\pi_A(t,\omega_c) - \sum_{t\in S_B(s)} p_B(t)\pi_A(t,\omega) \tag{8}$$

Two things to notice here: (i) if $\omega = \omega_c$ then the expected difference is 0; and (ii) if $\omega = \omega_d$ and payoff information is only privately known (and so if $s \in \mathsf{K_B}(P_B) \cap \overline{\mathsf{K_B}(P_A)}$) then the expected difference is 0.

Next, (and again, as is standard) we assume that B has an all-else-equal guilt parameter γ (where $\gamma \geq 0$). We further assume that her all-things-considered guilt in a state s for choosing an act that leads to ω ($\gamma(s,\omega)$) is her γ-scaled estimate of A's loss ($\delta_B(s,\omega)$) weighted by her degree of belief in the presence or absence of plausible deniability:

$$\gamma(s,\omega) = \max\{p_B(\mathsf{K_A}\mathsf{K_B}(P_A))\gamma\delta_B(s,\omega),\, 1 - p_B(\mathsf{K_A}\mathsf{K_B}(P_A))\gamma_0\delta_B(s,\omega)\} \tag{9}$$

where $\gamma_0 = r\gamma$ for $r \in (0,1)$. Intuitively, this is how much of B's guilt parameter is driven by her first-order or "pure" guilt that isn't sensitive her higher-order

[6] While we opt for the guilt-talk here, it is important to keep in mind that we are trying to be neutral here between theories of guilt aversion and reciprocity theories based on implicit contracts. They both ought to make room for higher-order information to matter to motivation and we are aiming for a neutral way to formalize that in order to derive predictions.

information.[7]

Finally, given a state s, B's guilt-adjusted expected utility of a choice is simply an adjusted expected value, where the adjustment is her all-things-considered guilt.

$$GEU(\mathsf{Coop}) = \sum_{s \in S_B(s^*)} p_B(s)(\pi_B(s, \omega_c) - \gamma(s, \omega_c)) \tag{10}$$

$$GEU(\mathsf{Defect}) = \sum_{s \in S_B(s^*)} p_B(s)(\pi_B(s, \omega_d) - \gamma(s, \omega_d)) \tag{11}$$

To reiterate: our purpose in casting more or less familiar models in a framework that allows us to explicitly model uncertainty about (higher-order) information about payoff possibilities is to show how, given such a framework, we can generate predictions by varying distributions of that information. We can now do that for models in the strong reciprocity/guilt aversion family.

So consider, first, the distribution of information outlined in PRIVATE: in s^*, B only knows her own payoff possibilities. Thus $\delta(s^*, \omega_d) = 0$ and so $\gamma(s^*, \omega_d) = 0$. Hence:

$$GEU(\mathsf{Coop}) = \sum_{s \in S_B(s^*)} p_B(s)\pi_B(s, \omega_c) \tag{12}$$

$$GEU(\mathsf{Defect}) = \sum_{s \in S_B(s^*)} p_B(s)\pi_B(s, \omega_d) \tag{13}$$

But since every s compatible with B's information is one in which $\pi_B(s, \omega_c) = 25$ and $\pi_B(s, \omega_d) = 40$, a rational B will choose Defect. This is the right result: higher-order belief dependent motivations should be idle in the absence of the relevant higher-order information.

Now consider things at the other extreme: information about payoff possibilities for A and B is common knowledge between them as in COMMON. Then there is no uncertainty about the payoff possibilities each faces, higher-order or otherwise. Hence since B knows P_A, $\delta(s^*, \omega_d) = 15$. And since this is common knowledge, $p_B(\mathsf{K_A K_B}(P_A)) = 1$. From which it follows that $\gamma(s^*, \omega_d) = \gamma 15$. Hence, for modest values of γ, $GEU(\mathsf{Coop}) > GEU(\mathsf{Defect})$. And so rational

[7] We assume that r is generally small. You can think of this as leaving open the possibility that some agents can have at least a small twinge of guilt even when they enjoy plausible deniability and that if the source of the twinge accounts for enough of γ or if $\delta_B(s, \omega)$ is large enough relative to $\delta_B(s, \omega)$, then such agents may cooperate.

but modestly guilt prone B's will choose Coop. This, again, is the right result: we have derived that the standard higher-order belief dependent motivations should be active in the presence of the relevant higher-order information.

Now take the interesting intermediate case where payoff information is distributed as in ASYMMETRIC. Here B knows the relevant payoff possibilities for both A and B but she also knows that A does not know that B knows the possibilities A faces. Thus, since B knows P_A, $\delta(s^*, \omega_d) = 15$. But since B knows that A doesn't know this, $p_B(\mathsf{K_A K_B}(P_A)) = 0$. Hence $\gamma(s^*, \omega_d) = r\gamma(15)$ and so there are modest values of γ such that $GEU(\mathsf{Coop}) < GEU(\mathsf{Defect})$.

That gives us a concrete hypothesis: a higher-order belief motivated B will choose Defect in an environment like PRIVATE and she may cooperate in an environment like COMMON (depending on just what her γ value is). But her behavior in COMMON doesn't determine her choice behavior in an environment like ASYMMETRIC: such a B may cooperate in COMMON but defect in ASYMMETRIC. Put another way: if the population includes a substantial number of homo reciprocans, then the rate of defection in COMMON should be lower than the rate of defection in ASYMMETRIC which in turn should be lower than the rate of defection in PRIVATE.

4 Design and hypotheses

Our design is straightforward: a simple between-subject design consisting of one-shot trust games played between anonymously paired individuals where the only variable is the distribution of payoff information (PRIVATE, COMMON, and ASYMMETRIC).

Our main hypothesis is that the well-attested cooperative play by principals in environments like the trust game depends, in part, on how higher-order information is distributed among the agents. Limiting that information can crowd out cooperation and crowd in defection. For ease of reference, we summarize our main hypothesis here.

Hypothesis. Defection increases as higher-order payoff information becomes distributed in ways further removed from COMMON. In particular:

1. There is more defection under ASYMMETRIC than COMMON.
2. There is more defection under PRIVATE than ASYMMETRIC.

If you prefer to think about these as predictions about specific models, then you can frame the hypothesis this way:

Hypothesis. The population is diverse with respect to types: opportunistic/narrow maximizing and fairness/inequity averse types do not exhaust the population. Some agents are motivated by higher-order beliefs about payoff information.

We expect to find confirming evidence for our hypotheses. We now turn to describing our procedures and reporting our results.

5 Experimental Procedures

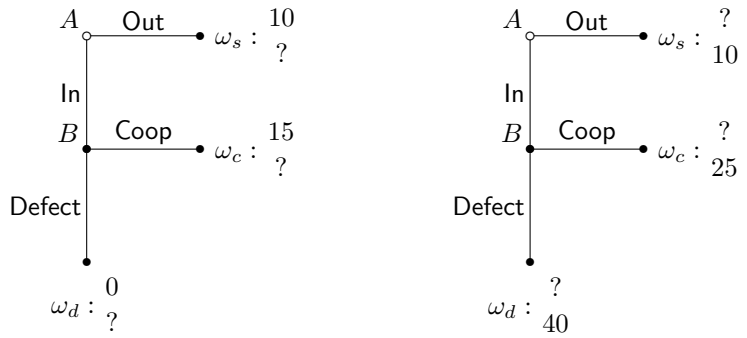
The subjects played the trust game once and only once, and this fact was common information prior to the session beginning.[8] Subjects were randomly assigned to the role of "Decision Maker 1" (player A in the game) or "Decision Maker 2" (player B); A's and B's were kept separate for the entire experiment using two rooms. Once an even number of subjects had arrived, instructions were handed out, and then read aloud. Subjects were allowed to ask questions individually. When there were no additional questions, the experiment began. All subjects received a large envelope which contained two smaller envelopes: one had the decision sheet and the other had a short demographic survey to be completed after all subjects had made their decisions. Subjects were asked to remove the decision sheet, record their move of right or down with an arrow, place the sheet back in the envelope, and drop the contents in a box at the back of the lab.[9]

The trust game was implemented using the strategy method: Bs were asked to assume that A had selected down and make their choice of "Right" (Coop) or "Down" (Defect) accordingly. This method allows us to collect data from all Bs, regardless of whether their matched A chooses In.[10]

[8] All sessions were run in the Robert Zajonc's Laboratory at the University of Michigan's Institute for Social Research. Subjects who had participated in similar experiments were excluded from recruitment.

[9] Data are available from the authors upon request. Most sessions had 20 subjects and took less than 1 hour to complete. Each subject received $5 for showing up on time. Average earnings (excluding the show-up payment) were $9.17 for As and $17.20 for Bs, and varied from $0 to $40.

[10] Some evidence exists that trust games implemented in this manner generate differences in behavior (McCabe, Smith & LePore, 2000; Coricelli, Morales & Mahlstedt, 2006; Solnick, 2007). Since we are mainly interested changes in behavior across information conditions, we can save those discussions for another day.



(a) A's Information (b) B's Information

Figure 2: Privatized Trust Game

The only difference between treatments was the description in the instructions of the information available to the players about potential payoffs. In COMMON, as is usual, subjects saw the full game tree and payoff possibilities available. Instructions were read aloud.

In PRIVATE, subjects saw privatized game trees (Figure 2: instructions for A show question marks that obscure the payoffs for B at terminal nodes and explain that these question marks hide B's payoff; and instructions for B show question marks that obscure the payoffs for A at terminal nodes and explain that these question marks hide A's payoff.

In ASYMMETRIC, instructions for A are as in PRIVATE; but B sees the full unobscured game tree (as in COMMON) and sees the game tree and discussion as it appears in A's instructions.

In each treatment subjects completed a post-instruction quiz that was checked for accuracy to ensure that they understood the information distribution: A had to demonstrate an understanding of the relevant information distribution; and B had to answer the higher-order question about what A would know under the relevant information distribution. Subjects showed no difficulty with the quiz. At the completion of the experiment, subjects filled out a short

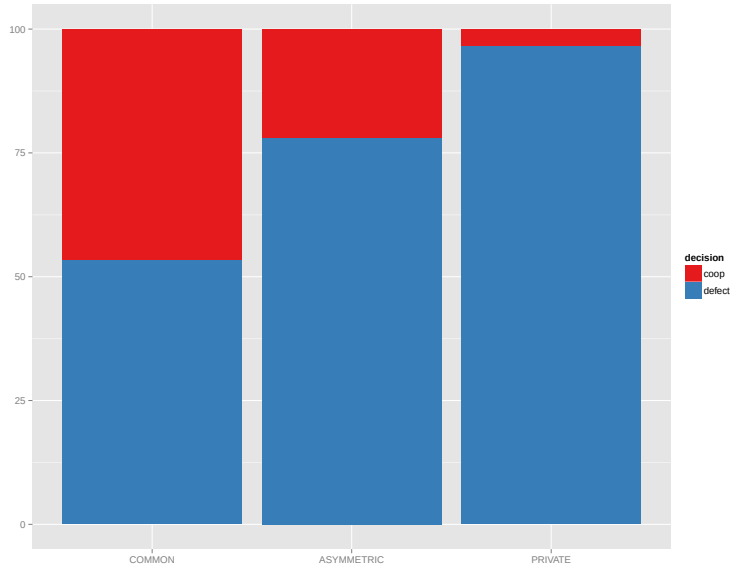


Figure 3: Rates of Defection (Blue) and Cooperation (Red) by B

demographic survey.[11]

6 Main results

Our main result, as we will see, confirms our main hypothesis that shifting higher-order information about payoffs away from common knowledge opens the door for more defection: subjects do in fact exploit plausible deniability.

Result.

1. $\Pr(\textsf{Defect}|\textsf{COMMON}) \ll \Pr(\textsf{Defect}|\textsf{ASYMMETRIC})$
2. $\Pr(\textsf{Defect}|\textsf{ASYMMETRIC}) \ll \Pr(\textsf{Defect}|\textsf{PRIVATE})$

We now discuss how the data support this conclusion.

First we simply note that the proportion of Bs choosing Defect differs across treatments, increasing from COMMON to ASYMMETRIC to PRIVATE (all

[11] The ease with which subjects understand and track higher-order information might seem somewhat surprising. But this coheres well with the success of reality TV shows that rely heavily on manufactured drama about who knows what about who does and doesn't know some tidbit. The quiz and survey are available from the authors upon request.

Defection Rates By Treatment			
	COMMON	ASYMMETRIC	PRIVATE
Bs Choosing Defect (%)	54.49	88.00	96.67
# of Pairs	43	50	30
Difference in Proportions Tests			
Pr(Defect|PRIVATE) $\gg$ Pr(Defect|ASYMMETRIC)		Yes*	
Pr(Defect|ASYMMETRIC) $\gg$ Pr(Defect|COMMON)		Yes**	

p-values: $^{*} \leq 0.05$, $^{**} \leq .01$, $^{***} \leq .001$

Table 1: Defection Rates

inequalities strict, and differences significant and large). The rate of defection in ASYMMETRIC is statistically higher than in COMMON ($p = 0.0124$) and the rate of defection in PRIVATE is statistically higher than in ASYMMETRIC ($p = 0.0236$).[12] This is depicted graphically in Figure 3. Note that defection rates begin in COMMON at modest levels (roughly similar to defection rates in other (strategy method) trust game experiments) and steeply rise in ASYMMETRIC and rise again (as expected) in PRIVATE. From the point of view of plausible deniability, it is the high frequency of defection in ASYMMETRIC that is intermediate between the levels in COMMON and PRIVATE that is most relevant. This is summarized in Table 1.

Next, we estimate the following logistic model:

$$Defect = \beta_0 + \beta_1(Asymmetric) + \beta_2(Private) + \beta_3(Gender) + \epsilon \tag{14}$$

where $Defect = 1$ if B chose Defect (0 otherwise), $Gender = 1$ if B identified as female (0 otherwise), and $Asymmetric$ and $Private$ are treatment dummy variables.

The results of the logistic regression analysis are summarized in Table 2.[13] The first column reports the logistic regression coefficients, the second column reports the standard errors, and the third column reports the odds ratio. In

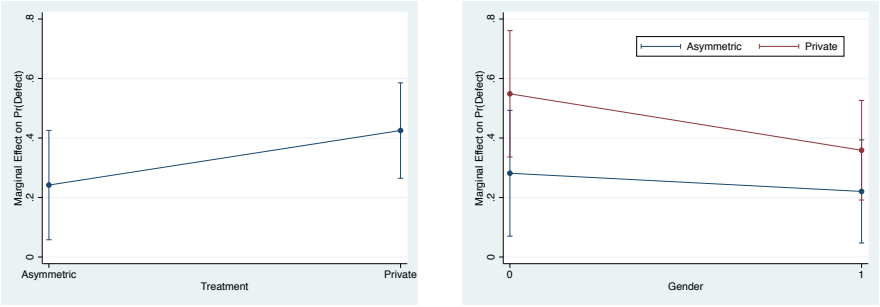
[12] The results reported here are based on difference in proportion tests, unless otherwise noted.

[13] We estimated the model with controls for other demographic characteristics, such as age, major, family income, etc. The regression results are not reported here as none of the variables were significant in and yielded nearly identical results to the regressions without the demographic controls. All are available upon request.

Predicting Defection			
	β	SE β	e^{β}
Asymmetric	1.16**	0.47	3.20**
Private	3.25**	1.07	25.69**
Gender	0.90*	0.46	2.47*
Constant	−0.43	0.43	0.65
χ^2		24.04	
pseudo R^2		0.1705	

p-values: * ≤ 0.05, ** $\leq .01$, *** $\leq .001$; e^{β} is the odds ratio

Table 2: Summary of Logistic Regression Analysis for Variables Predicting Bs Decision to Defect



(a) By Treatment (b) By Treatment × Gender

Figure 4: Average Marginal Effects

support of our main result, note that: (i) the odds of defection by Bs in ASYMMETRIC are more than 3 times greater than those under COMMON ($p = 0.01$); and (ii) the odds of defection by Bs in PRIVATE are more than 25 times greater than those under COMMON ($p = 0.002$). Additionally, the odds of defection by female Bs is about 2.5 times higher than by male Bs ($p = 0.05$).

Finally, we extract average marginal effects based on this model.[14] This is recorded in Figure 4.

We treat COMMON as our baseline condition; hence the marginal effect

[14]For a discussion of average marginal effects (and a practical guide to calculating them in R) see Leeper 2017.

on $\Pr(\textsf{Defect}|\textsf{COMMON}) = 0$. So deviation from 0 represents a marginal effect on the probability of defection. Note in Figure 4a that the probability that B chooses Defect in ASYMMETRIC is $\approx .24$ higher than in COMMON ($p = 0.01$). And the effect is even more pronounced in PRIVATE: an average marginal effect on $\Pr(\textsf{Defect})$ of $\approx .43$ ($p = 0.0001$). These effects are independent of gender, as shown in Figure 4b: in ASYMMETRIC, there is an average marginal effect for both males and females ($\approx .28$ ($p = 0.009$) and $\approx .22$ ($p = 0.01$), respectively); in PRIVATE, there is also an average marginal effect for both males and females ($\approx .55$ ($p = 0.0001$) and $\approx .36$ ($p = 0.0001$), respectively). This is strong evidence that defection rates increase as higher-order payoff information moves away from common knowledge.

7 Conclusion

Our primary interest was in probing B's motivation when she cooperates by structuring the information she has in ways that might tempt her to defect. The results we report here are stark. When payoff information is common knowledge, subjects manage to reach Pareto superior cooperative outcomes. This is, of course, well attested. What we see here is that when that information is less than common knowledge, defection by B gets crowded back in. This is less surprising when that information is fully private, but more surprising when B has plausible deniability. Otherwise trustworthy types do seem to be motivated in part by their higher-order beliefs about the distribution of monetary payoffs. When B knows the payoffs A faces and knows that A does not know that she knows this, B can be tempted to pursue her dominant action.

While probing B's motivation was our main interest, we note two further (and related) conclusions. First, we have shown that there is a theory-neutral framework for representing agents' information about payoff information and that that framework is expressive enough to re-state other familiar theories of B's choice behavior in it. Doing that across the information conditions here lead directly to predictions for those models. Thus, second, the results here offer a novel window into testing theories of off-equilibrium cooperative behavior in situations like the trust game. In particular, while our results are broadly consistent with the strong reciprocity theories (broadly construed), they are not consistent with the predictions of fairness/inequality aversion models (again, broadly construed). Put another way: our results show that mere manipulation of higher-order payoff information, as opposed to manipulations

of the strategic environment itself, reveals the existence of strong reciprocity types in the population. Some strong reciprocators exploit plausible deniability.

Bibliography

Attanasi, Giuseppe, Pierpaolo Battigalli & Elena Manzoni. 2016. Incomplete-information models of guilt aversion in the trust game. Managment Science 62(3). 648–667.

Battigalli, Pierpaolo & Martin Dufwenberg. 2006. Guilt in games. American Economic Review 97(2). 170–176.

Berg, Joyce, John Dickhaut & Kevin McCabe. 1995. Trust, reciprocity, and social history. Games and Economic Behavior 10. 122–142.

Bohnet, Iris & Richard Zeckhauser. 2004. Trust, risk and betrayal. Journal of Economic Behavior & Organization 55(4). 467–484.

Camerer, Colin F. 2003. Behavioral game theory. Princeton University Press.

Charness, Gary & Martin Dufwenberg. 2006. Promises and partnership. Econometrica 74(6). 1579–1601.

Coricelli, Giorgio, Luis Gonzalez Morales & Amelie Mahlstedt. 2006. The investment game with asymmetric information. Metroeconomica 57(1). 13–30.

Dana, Jason, Roberto A. Weber & Jason Xi Kuang. 2006. Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. Economic Theory 33(1). 67–80. doi:10.1007/s00199-006-0153-z. URL http://dx.doi.org/10.1007/s00199-006-0153-z.

Falk, Armin, Ernst Fehr & Urs Fischbacher. 2008. Testing theories of fairness – intentions matter. Games and Economic Behavior 62(1). 287–303.

Fehr, Ernst & Klaus M. Schmidt. 1999. A theory of fairness, competition, and cooperation. The Quarterly Journal of Economics 114(3). 817–868.

Hoffman, Elizabeth, Kevin A. McCabe & Vernon L. Smith. 1998. Behavioral foundations of reciprocity: experimental economics and evolutionary game theory. Economic Inquiry 36. 335–352.

Johnson, Noel D. & Alexandra A. Mislin. 2011. Trust games: A meta-analysis. Journal of Economic Psychology 32(5). 865–889. doi:10.1016/j.joep.2011.05.007. URL http://dx.doi.org/10.1016/j.joep.2011.05.007.

Leeper, Thomas J. 2017. Interpreting regression results using average marginal effects with R's margins. Tech. rep. URL https://cran.r-project.org/web/packages/margins/index.html.

McCabe, Kevin, Mary Rigdon & Vernon L. Smith. 2002. Cooperation in single play, two-person extensive form games between anonymously matched players. In Rami Zwick & Amnon Rapoport (eds.), Experimental business research, chap. 3, 49–67. Boston, MA: Kluwer Academic Publishers.

McCabe, Kevin A., Mary L. Rigdon & Vernon L. Smith. 2003. Positive reciprocity and intentions in trust games. Journal of Economic Behavior & Organization 52(2). 267–275.

McCabe, Kevin A. & Vernon L. Smith. 2000. A comparison of naive and sophisticated subject behavior with game theoretic predictions. Proceedings of the National Academy of Sciences 97(7). 3777–3781.

McCabe, Kevin A. & Vernon L. Smith. 2001. Goodwill accounting and the process of exchange. In Gerd Gigerenzer & Reinhard Selten (eds.), Bounded rationality: The adaptive toolbox, 319—342. MIT Press.

McCabe, Kevin A., Vernon L. Smith & Michael LePore. 2000. Intentionality detection and 'mindreading': Why does game form matter? Proceedings of the National Academy of Sciences 97(8). 4404–4409.

Ortmann, Andreas, John Fitzgerald & Carl Boeing. 2000. Trust, reciprocity, and social history: A re-examination. Experimental Economics 3(1). 81–100.

Smith, Vernon L. 2004. Human nature: An economic perspective. Daedalus 133(4). 67–76.

Smith, Vernon L. 2008. Rationality in economics: Constructivist and ecological forms. New York, NY: Cambridge University Press.

Smith, Vernon L. & Bart Wilson. 2014. Fair and impartial spectators in experimental economic behavior. Review of Behavioral Economics 1(1). 1–26. doi:10.1561/105.00000001. URL http://dx.doi.org/10.1561/105.00000001.

Smith, Vernon L. & Bart J. Wilson. 2017. Sentiments, conduct, and trust in the laboratory. Social Philosophy and Policy 34(01). 25–55. doi:10.1017/s0265052517000024. URL http://dx.doi.org/10.1017/S0265052517000024.

Snijders, Chris & Gideon Keren. 2001. Do you trust? Whom do you trust? When do you trust? In Shane R. Thye, Edward J. Lawler, Michael W. Macy & Henry A. Walker (eds.), Advances in group processes, vol. 18, 129–160. JAI, Elsevier Science.

Solnick, Sara J. 2007. Cash and alternate methods of accounting in an experimental game. Journal of Economic Behavior & Organization 62(2). 316–321.