On Blame-freeness and Reciprocity: an Experimental Study

Mariana Blanco, Bogachan Celen and Andrew Schotter

June 2017

Motivation

- In recent years, both in theoretical and experimental literatures, people have investigated the idea of reciprocal behavior:

Rabin (1993)
Falk and Fischbacher (2006)
Dufwenberg and Kirschteiger (2004)
Battigalli & Dufwenberg (2009)
Levine (1998)
Segal and Sobel (2007, 2008)
Sobel (2005) survey

What is Kindness?

- Reciprocity is the idea that people are willing to reward nice or kind acts and to punish unkind ones.
- This type of reciprocity, can be seen in many situations.
- If reciprocity means returning kindness with rewards and unkindness with punishments, however, it seems as if we have to define what kindness means.
- In this paper we propose a new definition of kindness called "blame-freeness".
- Test it using a simple experiment.

Blame Freeness: A Definition

- *Our notion of blame states that in judging whether player i has been kind or unkind to player j, player j would have to put himself in the strategic position of player i and ask himself how he would have acted under identical circumstances.*
- *If j would have acted in a worse manner than i acted, then we say that j does not blame i for his behavior. If, however, j would have been nicer than i was, then we say that "j blames i" for his actions (i's actions were blameworthy.)*

Properties of Blame Freeness

- Blame worthiness is only a necessary condition for punishment.
- The important point is that people use their own personal standards (how they would have behaved) to judge the actions of others and not some external norm like equity etc.
- This idea furnishes the predictions that we test in our experiments.

Properties of Blame Freeness (continued)

- This view of kindness is a **process oriented**, **endogenous view** that is **context or institution dependent**.
- Few of the extant theories of fairness or reciprocity share all of these the features.

Example

The Ultimatum Game and End-State Theories

- Consider the Ultimatum Game played between a Proposer, P, and a Receiver R.
- According to Fehr and Schmidt, Bolton-Ockenfels and any other end-state theory, an allocation (x_p, x_r) is rejected if

$$U_r(x_p, x_r) < U_r(0, 0).$$

- According to Blame the decision to reject depends on what he you would have done in the position of the Proposer.
- Let (x_p^*, x_r^*) be the allocation you would have made as a Proposer.
- If the current $U_r(x_p, x_r) > U_r(x_p^*, x_r^*)$, then you accept.
- If the current $U_r(x_p, x_r) < U_r(x_p^*, x_r^*)$, then you reject - - maybe.
- So you compare $U_r(x_p, x_r)$ to $U_r(x_p^*, x_r^*)$ and not $U_r(x_p, x_r)$ to $U_r(0, 0)$.

Players, actions, histories and preferences.

- Consider a sequential game consisting of two players $i = 1, 2$.
- $\mathcal{H}$ denotes the set of all histories.
- When it is player i's turn, he takes an action from the set of actions $A_i(h)$ that is available to him at $h \in \mathcal{H}$.
- A history h is terminal if $A_i(h) = \varnothing$ for all i. We refer to a terminal history as an outcome and denote the set of all outcomes by H.
- Each outcome h is associated with a material payoff for each player.
- The function $\pi_i : H \rightarrow R$ determines player i's material payoff $\pi_i(h)$ at outcome h.

Strategies

- Player i's *strategy* is a function $\sigma_i : \mathcal{H} \backslash H \rightarrow A_i(h)$ that determines an action at each non-terminal history for player i.
- Note that each strategy profile $\sigma = (\sigma_1, \sigma_2)$ induces a unique outcome $h \in H$. For a given σ, we write $h_\sigma \in H$ to denote the outcome induced by σ.

Putting Oneself in Another's Position

- Our definition of blame revolves around the comparison of two entities:

1. What a player would do if he were in other player's position (and hence what he thinks his payoff would be when he plays against himself).
2. What he thinks his opponent will do when he plays against him and hence what he thinks his payoff will be when he faces his opponent.

- We follow the Psychological Game literature initiated by Geanakoplos et al. (1989) and modified by Dufwenberg and Kirchsteiger (2004) and Battigalli and Dufwenberg (2009) .

Putting oneself in Another's Position: Beliefs.....

- Our definition of blame requires reference to two levels of beliefs:

1) player i's belief about player j's strategy, $\hat{\sigma}_{ij}$

2) player i's belief about player j's belief about i's strategy, $\hat{\sigma}_{iji}$

- Denote player i's beliefs by $\mu_i := (\hat{\sigma}_{ij}, \hat{\sigma}_{iji})$, and profile of players' beliefs by $\mu := (\mu_1, \mu_2)$.

Putting oneself in Another's Position: Strategy

- Since what player i would do in his opponent's position is key to our discussion we will refer to player i's strategy if he were in player j's position by σ_{ij}.
- We denote the strategies of player i by $s_i := (\sigma_i, \sigma_{ij})$ and a profile of strategies by $s := (s_1, s_2)$.
- We assume that there is an underlying preference structure behind the strategy σ_{ij}.
- More precisely, we assume that player i is endowed with preferences that he would have if he were in player j's position and that the preferences are represented by a utility function u_{ij}- - to be defined later.

What I Think My Opponent Will Give Me

- When a profile $\hat{\sigma}$ consists of beliefs $\hat{\sigma} = (\hat{\sigma}_{iji}, \hat{\sigma}_{ij})$ it induces a unique outcome which we by $h_{\hat{\sigma}}$.

– This indicates what I think the outcome will be when my opponent plays against me given what I think he thinks I will do.
– We write $\pi(h_{\hat{\sigma}})$ to denote the expected material payoff from the profile $\hat{\sigma}$.

What I Would Give Me in His position

- When i puts himself in j's position, his belief about i's (his own) strategy would be $\hat{\sigma}_{iji}$.
- Given this belief, if i were in j's position, he would play σ_{ij}which is a best response to $\hat{\sigma}_{iji}$.
- Hence, if i were in j's position, i would create a material payoff of $\pi_i(h_{(\hat{\sigma}_{iji},\sigma_{ij})})$ for himself. This is how kind he would be to himself if he were in j's position.

Blame....

- So we have two payoffs for i:
- What he thinks his payoff will be when playing against j given beliefs: $\pi_i(h_{\hat{\sigma}_{iji},\hat{\sigma}_{ij}})$.
- What he thinks his payoff would be if he played against himself in j's position given his beliefs: $\pi_i(h_{(\hat{\sigma}_{iji},\sigma_{ij})})$.
- The difference between his expected material payoff from of player j (when j plays against him) and the expected material payoff he would expect to get when he is in j's position playing against himself, is the source of *blame*.

Definition of Blame....

Definition

Given the strategy and belief profile (s_i, μ_i)*, player i is said to blame player j if* $\delta_i(s_i, \mu_i) := \pi_i(h_{(\hat{\sigma iji}, \sigma ij)}) - \pi_i(h_{(\hat{\sigma iji}, \hat{\sigma ij})}) > 0.$

- **Statement of player i:**

"*I blame player j because the material payoff which I believe he expects to give me if I play against him is less than my expected material payoff if I played against myself. In other words, if I was in his position I would be nicer to a player in my position than I expect him to be to me.*"

- j is being more unkind to me than I would be to me in his position.

Utility Function.....

Player i's utility function

- We argue that blame affects a player's altruism towards his opponent and incorporate it in the preferences as follows.

$$u_i(s,\mu) := v_i(\pi_i(h_\sigma)) + \beta_i(\delta_i(s_i,\mu_i))\pi_j(h_\sigma),$$

where β_i is non-increasing in blame $\delta_i(s_i,\mu_i)$.

- The utility function u_i indicates that player i's utility is determined by the sum of utility v_i derived from his material payoff and a proportion of player j's material payoff.
- The term β_i is the weight attached to other player's material payoff and it is determined by how much player i blames player j.

Utility Functions ....

Player i's utility function in j's position

- Player i's preferences in player j's position is represented by a function u_{ij} defined as

$$u_{ij}(s,\mu) := v_i(\pi_j(h_{(\hat{\sigma}_{iji},\sigma_{ij})}) + \beta_i(0)\pi_i(h_{(\hat{\sigma}_{iji},\sigma_{ij})})).$$

- That is, when he considers himself in player j's position he adopts the belief $\hat{\sigma}_{iji}$ about his (player i's) strategy and plays σ_{ij}.

Blame evolves in the course of a game....

- Given our set up we can now define an equilibrium as a set of beliefs and strategies that are consistent.
- However, in our game players start out with initial beliefs at the root of the game tree $\varnothing$.
- As the game evolves they may find themselves at a node they did not expect and must update their beliefs about the strategies being used by their opponent.
- In such a case, the player revises his beliefs to be consistent with the node reached.
- Note that as these beliefs are updated the payoffs at the terminal nodes are also changed since different beliefs imply different amounts of blame which change payoffs.

Equilibrium ......

Definition: *Sequential Blame equilibrium*

The profile (s, μ) is a SBE if for $i, j \in \{1, 2\}$ and for each history $h^* \in \mathcal{H}$, the following holds:

1. $\sigma_i \in argmax_{\sigma_i} u_i^{h*}((\sigma_i, \sigma_j), \mu_i)$,
2. $\sigma_{ij} \in argmax_{\sigma_{ij}} u_{ij}^{h*}((\sigma_i, \sigma_{ij}), \mu_i)$
3. $\hat{\sigma}_{ij} = \sigma_j, and \hat{\sigma}_{iji} = \sigma_i$.

- This specifies:

1. sequential rationality for each player.
2. sequential rationality for players put in other's positions
3. Consistency of beliefs - - self confirmation.

Testing Blame in Public Goods Context

- To test Blame experimentally need two features:

1. It must be possible to have a player play in both his and his opponent's position so we can observe his actions. .
2. There must be room for reciprocation or punishment.
3. These two features exist in public goods games with punishment since those games are symmetric and involve punishment.

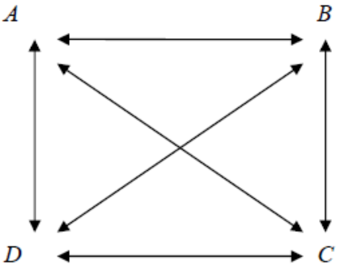
Public goods with punishment

- Subjects play a 4-player game. Each subject is endowed with y tokens.
- In stage 1 subjects simultaneously choose their contribution $g_i \in [0, y]$.
- Each subject i's payoff is $\pi_i := y - g_i + \alpha G$ where $\alpha \in (0, 1)$ and $G = \sum_i g_i$.
- At stage 2, a subject is allowed to *punish.*
- Punishment is costly $c \in (0, 1)$. If subject i punishes subject j by reducing subject j's payoff by p_i^j,i, he incurs a cost of cp_i^j
- If the subject's utility is an increasing function of $\pi_i = y - g_i + \alpha G$ then in the subgame perfect equilibrium of the game $p_i^j = 0$ and $g_i = 0$ for all i, j.

Networks and Punishments in Public Goods games

- Standard public goods games involving punishment involve subjects connected in a complete network.

Complete Network



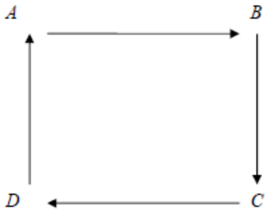
- Typical conclusion is that those who contribute less than some exogenous norm (mean) get punished.
- De Quervain et al. (2004) Fehr and Gächter (2000) to mention just two papers

Problem - - Identifying Motives

- Data shows that people who contribute below mean get published more - - but not why.
- Because of compete network we can not identify motives for punishment.
- Many theories would suggest punishing those who contribute least - - they are the worst offenders in all theories.
- What if you contribute above the mean but less than me?
- What if you contribute below the mean but more than me?
- Changing network could help.

Public goods with punishment - - Directed Circle

- *Carpenter, Kariv, and Schotter (2012), "Network architecture and mutual monitoring in public goods experiments", Review of Economic Design, 2012.*
- You only get to punish the person you observe and he does not punish you.



Blame

- Blame suggests that people punish those that contribute less than they do whether that is above or below the group mean.
- You judge others by how you behaved not by some exogenous equity norm.
- If you contributed way below the mean why punish another who gave more than you but also below the mean.
- Correct way to do this is to use the Circle Network $A \rightarrow B \rightarrow C \rightarrow D \rightarrow A$
- You are also told the mean contribution

Regression

- p_{ci} = public good contribution of subject i ,
- p_{c-i} = the target's contribution,
- $\Delta^{+}_{other} = pc_i - pc_{-i}$ *if* $pc_i - pc_{-i} > 0$ and 0 otherwise
- $\Delta^{-}_{other} = pc_i - pc_{-i}$ *if* $pc_i - pc_{-i} < 0$ and 0 otherwise.
- $pc_{-i} - m$ is the difference between a target's contribution and the mean.
- m is the mean contribution of the group.

$$Pr(punishment) = \alpha + \beta_1 \Delta^{+}_{other} + \beta_2 \Delta^{-}_{other} + \beta_3 (pc_{-i} - m) + \beta_4 (m) + \epsilon_i.$$

Results

Regression

- If the theory of blame is responsible for punishment behavior we would expect that coefficient β_1 would be positive and significant while all other coefficients should be insignificantly different from zero.
- All that should matter for punishment is whether a subject's contribution was more or less than the contribution of the person he monitored.
- The regression results are presented in Table 2.

Regression Results

TABLE 3: Punishment Behavior in CKS: Directed Circle

	Coefficient	Z	$p > \|Z\|$
Δ^{+}_{other}	.153	4.86	.000
	(.032)		
Δ^{-}_{other}	.046	1.45	.148
	(.032)		
$(pc_{-i} - m)$	-.038	-1.16	.246
	(.030)		
mean	-.014	-.44	.663
	(.033)		
constant	-1.422	-2.24	.025
	(.636)		

$N = 240$, Wald $\chi^2(4) = 45.92$, Pr $> \chi^2 = .000$.
Robust z-statistics are reported in parentheses (clustering at the subject level.)

- As we see, only the difference between one's own contribution and that of the target matters.
- The mean not insignificant.
- Note that if they could see all contributions they might punish only those

Other Takes on the Data: Comparing Complete and Directed Circle Networks

- Look at two-tiered decision: whether to punish and how much.
- Use hurdle regression.

1. Fit a logit regression on whether to punish or not (the binary choice on the extensive margin) using the same set of explanatory variables as used above.
2. Fit a Poisson regression of the punishment level (the discrete choice on the intensive margin) using the same set of explanatory variables.
3. Estimated via maximum likelihood method to jointly for both the Directed and the Complete networks.

TABLE 4: Hurdle Model Estimation

	(1)	(2)	(3)	(4)
	Directed		Complete	
Punishment Margin	Extensive	Intensive	Extensive	Intensive
$(pc_i - pc_{-i})^+$	$.212^{***}$	$.034^{*}$	$.054^{***}$	$.001$
	(.040)	(.018)	(.020)	(.010)
$(pc_i - pc_{-i})^-$	$-.007$	$-.041$	$.160^{***}$	$-.049^{***}$
	(.033)	(.027)	(.019)	(.018)
$(pc_i - \text{mean})$	$.040$	$-.019$	$.128^{***}$	$-.008$
	(.040)	(.019)	(.021)	(.014)
mean	$.004$	$-.010$	$-.018$	$-.076^{***}$
	(.043)	(.019)	(.018)	(.013)
constant	-1.885^{***}	1.607^{***}	$-.132$	2.055^{***}
	(.762)	(.427)	(.281)	(.196)
Observations	238	238	716	716

Robust standard errors in parentheses.
$^{***}\ p < .01,\ ^{**}\ p < .05,\ ^{*}\ p < .10,$

Directed Networks (columns (1) and (2)):

1. key driver of both the decision to punish and its magnitude is $(pc_i - pc_{-i})^+$.
2. $(pc_i - pc_{-i})^-$ not significant.
3. Any variable referring to the mean contribution of the group is not significant.

Complete Network (columns (3) and (4)):

1. $(pc_i - pc_{-i})^+$ significant for decision to punish not its level.
2. $(pc_i - pc_{-i})^-$ significant for both decision to punish and level. "anti-social punishment" people punished for contributing more.
3. Directed circle adds support to "revenge motive" - - in Complete networks you assume that big contributors punished you, punish them as revenge.
4. Can't do that in Directed Circle.
5. Note that mean is significant in Complete Network for intensity of punishment. Artifact of seeing all contributions - - punish the worst.
6. Missing counter factual in Complete Networks would subjects punish others whose contribution were above the mean but below theirs.
7. The Directed Network answers that.

Conclusions

- We have presented a theory of kindness and reciprocity based on the notion of blame.
- We have tried to show that such a notion is process-oriented, endogenous and context dependent.
- We have shown that the theory makes predictions distinct from those of other widely used theories.
- We have reported on the results of a simple public goods experiment to demonstrate that this notion of blame-based reciprocity does have some power to organize data.
- Finally, note that we are not saying that this notion should replace other notions of kindness or reciprocity but rather it can coexist with others in the population of people - - some may adhere to exogenous norms while others may evaluate the action of others in terms of their own internal code of ethics.
- Their own code of ethics could be shaped by existing norms so they