The One who Gives *Too Early*, Gives Twice: Why Does Cooperation Tend to Decay?

Elias L. Khalil[*]

ABSTRACT

How to explain cooperation in laboratory finite public good experiments? For traditional economics, humans are prone to free-ride and, hence, any cooperation must have risen as a result of confusion on the part of participants. But confusion cannot explain all the observed cooperation. For behavioral economics, cooperation should be expected since humans are motivated by "prosocial preferences." But then why would cooperation tend to decay in such experiments? This paper proposes a hypothesis that can easily explain *both* the authenticity of cooperation (not easily explained by traditional economics) and its eventual decay (not easily explained by behavioral economics). The hypothesis assumes that each participant is a perfect conditional cooperator, but starts to contribute less given the evolution of his or her belief about the contributions of other players in the game. Namely, the participant starts to develop the belief that others are contributing less than the agreed upon norm. This is paradoxical given that all others are also contributing their fare share, and each one also develops the belief that the others are generally "cheating." To explain such a miscommunication, the paper proposes that beliefs develop in light of the fact that income distribution is positively skewed (asymmetric). The positive skewness entails that, given the median income, a contribution that deviates (either up or down) from the expected contribution given the median income can arise either as a result of temperament (one wants to contribute early or late) or as a result of the true deviation of the contributor's income from the median income. The participant, if there is no asymmetric belief formation, should blame temperament with same likelihood irrespective of the direction of the deviation. But, for reasons explained in the paper, each participant tends to blame temperament more in the case of up-deviations than the case of down-deviations. So, each participant, when he or she observe that others are over-contributing or contributing too early, the participant would tend to expect them to have greater income than the median than it is truly the case. Such misplaced expectation leads to judgment that others are "cheating" on average, when it is not the case. Such misjudgment quickly leads to the decay of cooperation.

Keywords: Prosocial preferences; Coordination; *Homo economicus*; *Homo sociologicus*; Free-riding; Imperfect conditional cooperation; Beliefs; Unconditionally fair contribution; Fairness; Super-fairness; Positively Skewed Income Distribution; Restart Effect; Distorted Belief Formation; Seed Money and Leadership; Mediation

JEL Classification: D04

[*] Email: elias.khalil@buseco.monash.edu.au. Homepage: www.eliaskhalil.com. Department of Economics, Monash University, Clayton, VIC 3800, Australia. An earlier version received comments from Jason Aimone, Anmol Ratan, Aaron Nicholas, David Butler, Birendra Rai, participants of a workshop at Monash University, and especially the extensive comments of Ian McDonald. The usual caveat applies.

1. Introduction

T.E. Lawrence [1997], better known as “Lawrence of Arabia,” has provided a vivid account of murder and retribution in the Arabian Desert during the Arab revolt against the Ottoman Empire during War World I. One young man killed another young man from another tribe. A dispute arose: While both tribes agreed that the killer must be put to death, the tribe of the victim wanted to execute the punishment, to which the tribe of the perpetrator objected. Determined to maintain the peace between the two tribes before a major battle, Lawrence volunteers to be the executioner—a solution accepted by both tribes.

The need for a third party is simple. It is usually the case that the injured party demands a higher compensation than what the perpetrator party is ready to provide. If the perpetrator party acquiesced and allowed the victim’s tribe to be the executioner, the victim’s tribe might take it as a signal that it can demand a higher price, such as monetary compensation along with the execution. And if the perpetrator party objects to the demanded higher price, the situation might lead to the breakdown of cooperation in the sense of the rise of tribal hostilities, the disrespect of each other’s rights, and the open outbreak of blood feud or even war. Actually, the issue of “keeping the peace” is a social dilemma that involves cooperation. Social dilemmas are not limited to contributions to a fund to set up a public library or construct a road. They include the case of “keeping the peace”: In a finite game, each participant benefits from "teaching the other a lesson" which would facilitate the extraction of resources from the other. But if each one follows the dominant strategy of defection, it would lead to open hostility, a suboptimal Pareto outcome.

The relegation of decisions concerning punishment to a judge or a third part helps to avoid defection and the possibility of open hostility. If the punishment is specified by a judge or a third party, and the perpetrator party obeys the judge, it does not mean it agrees with the judge. It formally means that it is ready to obey the judge simply because of the horrible alternative. So, following the judge's order does not send a signal that the perpetrator is even more ready to pay a higher price. The use of third parties, and the function of the institution of magistrates in general, tend to put a stop to the dangerous evolution of beliefs that the perpetrator has to cough up more resources.

Put differently, when one party complies with the price demanded by the other party, gives too early to the demand of the other party, or pays more than what is judged as fair, it usually sends the wrong signal. It is potentially the wrong signal insofar as the other party uses it to update its belief that justifies raising the bar of what is fair even higher. Such raising would lead to the decay of cooperation. To avoid the distorted formation of beliefs, societies throughout history have set up diverse institutions, such as referees and judges and even unwritten norms. To understand such institutions, this paper proposes a model of the dynamics of belief formation that can account for the possibility of distorted belief formation.

2. The Cooperation Dilemma

The core thrust of the proposal is that distorted belief formation can explain the much documented decay of cooperation in public good games in laboratory experiments. The emphasis on belief formation is antidote to the emphasis on preferences in the majority of the literature. Irrespective of whether it is the standard neoclassical literature or the prosocial preferences literature, it largely explains the dynamics of cooperation ultimately in terms of the preferences of participants. The preferences can be self-interested or prosocial.

For the standard view at first approximation, cooperation in finite games in the laboratory should not have risen in the first place. The reason is the same as the one marshaled by Mancur Olson [1965] concerning the supply of public goods: There is an under-supply of public goods because of the free-riding problem or what is known in general social dilemmas. What is at the core of social dilemmas, which should rule out cooperation in finite games in the laboratory is that while it is optimal for the agent to defect, the social optimal is for all agents to cooperate.[1]

For the standard view, let us call it the *Homo economicus* view, each agent taking part in a finitely repeated game would not cooperate in the last round. As a result of backward induction, each agent would not cooperate even in the first round. The prediction of subgame perfect Nash equilibrium is defection in finitely repeated games even at the start of the game.

However, with some qualifications, the *Homo economicus* view can account for the rise of cooperation in finitely repeated games. For instance, David Kreps and Robert Wilson [1982]

[1] Social dilemmas take many forms such as the prisoners' dilemma, the public goods problem, the common resource problem, and the tragedy of the commons. In the laboratory, social dilemmas have been best expressed as Voluntary Contribution Mechanism (VCM) games [Isaac, McCue & Plott, 1985]. In general, a social dilemma arises when:

$$\pi_i = E_i - g_i + \beta \sum g_j$$
$$\text{Where } 1 > \beta > 1/n$$

where π_i is earning of i player, where i = 1, …, n; E_i the endowment and g_i the contribution of i player. If $\beta > 1$, and a player contributes \$1 and no one else contributes, the player would earn more by contributing even no one else is contributing. If $\beta < 1/n$, it does not pay for one to contribute \$1 even when others are contributing the same amount. Thus, to make cooperation a nontrivial issue, the above condition concerning β must hold. Then, the private marginal return of \$1 would be lower than \$1, while the social marginal return of \$1 would be higher than \$1.

introduce incomplete information in the finitely repeated game, and consequently show that cooperation arises. Further David Kreps *et al.* [1982] show how changing the belief structure can engender cooperation: If sufficient participants believe that a portion of the participants are irrational, then it pays to take their non-credible threats seriously and cooperate. Evolutionary game accounts [e.g., Frank ; 1988; Hirshleifer, 1987] demonstrate how the frequency of a trait, such as honesty, matters. If the frequency is sufficiently high, other self-interested players choose to play with honest participants, which helps the prorogation of the seemingly unfit trait.

But in laboratory experiments of finite public goods games, in which we witness cooperation, there is little uncertainty or beliefs about irrational actors. So, how can we explain cooperation? The advocates of the *Homo economicus* view face a challenge. One explanation is that agents can only examine or cognitively process the next four or five rounds and, hence, they start to defect towards the end of the rounds, i.e., once they are able to see the last round. However, through repetition and experience, participants should learn and should be able to start to defect at the start of the game. This does not seem to be the case, given the re-start effect discussed below.

Another explanation is that cooperation is the outcome of confusion [Binmore, 2006], which dissipates as a result of learning. Rigorous experiments in finite games have shown that confusion plays a role—but it cannot explain the whole range of cooperation [see Andreoni, 1995; Houser & Kurzban, 2002; Kurzban & Houser, 2005].

On the other hand, the prosocial economics approach,, which can be called the *Homo sociologicus* view, has no problem explaining the existence of genuine cooperation, i.e., cooperation in finite games with perfect information. It advances the notion that agents are motivated by "social preferences" aside from the usual self-interest motivation. Such social preferences entail that agents act on their commitment to advance the Pareto optimal outcome in public good experiments.

But, in turn, the *Homo sociologicus* view faces another kind of problem: It cannot explain why cooperation decays towards the end of the finite game—given that agents want to behave ethically? If agents are committed to ethical norms, why do we witness, systematically, defection towards the end of the game? One explanation offered by the *Homo sociologicus* view is that cooperation collapses probably because there are naïve participants who presume that most of the participants are honest and, hence, cooperative. And once they experience defection on the part of others, they are dismayed and start to defect, which sets in motion the decay of cooperation. But this explanation is *ad hoc*: For the explanation to be non-arbitrary, it must assume that there are equal numbers of non-naïve participants, i.e., skeptics. The skeptics, once they experience cooperation more than they expected, would become very cooperative. That is, while the naïve become less cooperative, the skeptics should become more cooperative—and both trends should offset each other. If so, there should not be a systematic tendency for cooperation to decay. The only way this explanation to work is to presume that the frequency of naïve agents exceeds the frequency of skeptic agents. But such a supposition would be *ad hoc* and, hence, can explain the decay of cooperation with great unease.

So, at hand, we have a puzzle, "the cooperation dilemma":

> **The Cooperation Dilemma:** While the *Homo sociologicus* view can explain the rise of cooperation, it cannot easily explain its decay. While the *Homo*

> *economicus* view can explain the decay of cooperation, it cannot easily explain its rise.

This paper offers a hypothesis that can easily explain both the rise and decay of cooperation. The hypothesis highlights the role of belief formation. This would put the hypothesis at a distance from the usual focus on preferences by both the standard *Homo economicus* view and the alternative *Homo sociologicus* view.

The proposed hypothesis on the importance of beliefs, is somewhat inspired by the approach of Fischbacher and Gächter [2010; see Gächter, 2006; Fischbacher *et al.*, 2001]. However, Fischbacher and Gächter set up a particular experimental design that eventually favors, again, the role of preferences, rather than beliefs, as the cause behind the decay of cooperation. In specific, Fischbacher and Gächter "find" that most participants are *imperfect* conditional cooperators: While most participants are ethical conditional cooperators, they are not perfectly ethical. While most of them are ready to conditionally cooperate with others, they also have a penchant for cheating a bit—which sets in motion the decay of cooperation.

The Appendix review in detail the sophisticated experimental set up of Fischbacher and Gächter and it shows its major weakness: Fischbacher and Gächter fail to control for the possibility that when agents reveal their "preferences" for a imperfect conditional cooperation—i.e., the taste for ethical behavior but with a bit of cheating—they are actually revealing a particular attitude shape by their past experience that is not controlled by the experimental design.

The proposed hypothesis springs from the methodological premise that it would be more effective to explain a phenomenon by avoiding the appeal to preferences, following the stipulation of Stigler & Becker [1977; see Khalil, 2008]. The proposed hypothesis, in particular, avoids the appeal to odd preferences such as imperfect conditional cooperators (i.e., ethical behavior that allows for some cheating). This paper shows, at least theoretically, how the decay of cooperation arises even when the participants in finite games are *fully* ethical, i.e., when they are *perfect* conditional cooperators.

The proposed hypothesis shows this possibility by relying on the formation of beliefs. The proposed hypothesis is at least a competing hypothesis to the ones that rely on preferences.

The hypothesis advances the thesis that belief formation becomes distorted even when agents follow Bayesian updating. The distortion leads to mis-coordination and such mis-coordination is the cause of the decay of cooperation. Others have examined the role of mis-coordination [e.g., Chwe, 2001; Chaudhuri, 2007; Young, 1993, 1996]. The issue of mis-coordination can somewhat be traced back to the theory of the firm advanced by Armen Alchian and Harold Demsetz [1972], where they argue that a team boss emerges because an egalitarian organization invites disputes over measurement of the contribution of each member. Each member tends to mis-measure, usually in one's favor, one's contribution to the output. The significance of the proposed hypothesis lies in changing the emphasis from monitoring, as a solution to solve the supposed problem of selfish preferences, to, instead, the improvement of communication. Poor communication might be more important than cheating as the source of the erosion of trust and the consequent decay of cooperation.

3. Assumptions of the Model

Let us examine a finite public good game with the following assumptions:

1. There are *n* agents (*n*≥2);
2. the rounds of play are finite and known;
3. contributions of agents are subject to shocks such as temporary temperament (mood swing) or health condition--where the shocks are *iid* and symmetrically distributed across each agent;
4. each agent is "conditionally fair," and assumes others to be likewise, in the sense that he or she is a perfect cooperator, but conditional on the cooperation of others. (That is, the agent is non-naïve: the agent is aware that others might cheat—but will not act without evidence);
5. each agent would initially contribute a fraction of one's actual income;
6. the fraction of income to be contributed is common knowledge—so, it is expected that participants will contribute unequal amounts in proportion to their unequal income;
7. each agent knows own income/capacity, which is assumed to be constant across the rounds of interaction, but does not know the particular income of each other participant and does not know the particular shock of each other participant;
8. agents share the same tacit hypothesis of the income distribution of the relevant society and such assessment is *tacit* in the sense that it is very weakly the subject of updating and, hence, assumed as fixed;

9. income distribution is positively skewed (skewed to the right);
10. the tacit hypothesis held by agents of income distribution is more skewed than what is actually the case.

Note, the assumption that the tacit hypothesis of income distribution is not subject of updating might be strong. But it is justifiable on the ground that the variance of the hypothesis is very low or at least much lower than the variance of the belief about a participant's particular income. As such, any change of the tacit hypothesis can be negligible and, hence, can be ignored at first approximation.

Concerning the assumption of temperament/shocks, it does not make preferences the motor of change. And, further, the shocks can be seen as states of the world, as the case with sickness or mood temperament, and hence they do not add any oddity to the structure of preferences.

Further, the assumption that all agents are perfect conditional co-operators (pCC) contradicts the standard experimental finding of free-riders. However, there is also a fraction of participants who are unconditional co-operators—i.e., who would not defect even when they witness others defecting. The fraction of unconditional co-operators may not totally offset the fraction of free-riders--but we have to examine the weight of each fraction in terms of contribution. The total contribution of each fraction might offset the other. In any case, even if free-riding is more prominent than unconditional cooperation, it may not be the sole culprit and, hence, it should be ignored if we want to investigate the possibility of distorted belief formation as at least a significant motor behind the decay of cooperation.

In addition, it is possible that players, as they play the actual game in laboratories, sense the decay and, hence, try to reverse it by heroic acts of contributions or act as leaders who are ready to contribute "seed money." This is probably the motive of the unconditional co-operators. But such acts come at secondary approximation to be worried about in the actual design of experiments. At first approximation, we can ignore remedies and institutions that agents undertake once they sense the natural course of belief formation which inevitably leads to the decay of cooperation.

4. The Model

In our model, we do not need to start with the standard maximization of utility since the agent is simply maximizing monetary return. In our model, the agent maximizes the monetary return subject to the honesty constraint of perfect conditional cooperation. Such a constraint specifies the decision rule of what to contribute, namely, a fraction of one's income, while taking into consideration the belief about the contribution of others.

Given the above assumptions, we can characterize the contribution (C) rule of agent i at period t, which is called "conditionally fair contribution,"

$$C_{i,t} = f(\alpha(y_i + s_t), \beta(C_{-i, t-1} - \check{C}_{-i}))$$

$$t = 1, \ldots, T$$

$$\text{Where } 0 < \alpha < 1;$$

$$\beta > 0; \text{ and } \beta(0) = 0$$

$$s_t = \begin{cases} s_H > 0 \text{ with probability } \pi_H \\ \end{cases}$$

$s_L < 0$ with probability π_L

$$s_H = -s_L$$
$$\pi_H = \pi_L = 50\%$$
$$E(y_i + s_t) = y_i$$

where y_i is income of agent i, Č unconditionally fair contribution (to be defined shortly), -i average other agents, and s_t shock that takes two values: s_H is the high income shock with probability π_H; s_L is the low income shock with probability π_L. And $E(y_i + s_t)$ is the expected income which is y_i, since the shocks are symmetrical in strength ($s_H = -s_L$) and equally probable ($\pi_H = \pi_L$).

The above contribution rule is defined as *conditionally fair contribution* (C):

Definition: conditionally fair contribution, as above,

$$C_{i,t} = f(\alpha(y + s_t), \beta(C_{-i,\, t-1} - \check{C}_{-i})),$$

$C_{i,t}$ is conditional because it depends on two variables: a) the actual income of the period (constant income and the current shock) of agent i; b) and gap between actual contribution of the previous period and the unconditionally fair contribution of agent -i.

Let us define now the *unconditionally fair contribution* (Č) of agent -i (or any agent):

Definition: unconditionally fair contribution

$$\check{C}_{-i} = f(\alpha E(y_{-i} + s_t)),$$

$$\text{that is, } \check{C}_{-i} = f(\alpha y_{-i})$$

So, the unconditionally fair contribution (Č) differs from the conditionally fair contribution (C) in three respects: i) unconditionally fair contribution is the same for all periods because what matters is expected income rather than actual income; ii) the β parameter is set to zero—since, bu definition, the contribution is not conditioned on the contribution of others; iii) unconditionally fair contribution of agent i is known to agent i, but it is known to agent –i only probabilistically.

In the first round, the agent's contribution (C) depends only on his or her income and shock given that there is no prior period to assess the contribution of the other agent upon which to formulate one's conditional contribution. So, in first round, each agent contributes, *on average*, the unconditionally fair contribution by definition. Note, though, the *actual* contribution of each agent diverges from the unconditionally fair contribution because of the idiosyncratic shock.

In second round, agent i takes into consideration the actual contribution of agent –i given the unconditional fair contribution, i.e., $(C_{-i,\, t-1} - \check{C}_{-i})$. Agent i may judge that agent –i has over-contributed (C^{o}), equally contributed (C^{e}), or under-contributed (C^{u}).

Definition: An over-contribution (C^{o}):

$$C_{-i,\, t-1} - \check{C}_{-i} > 0$$

Definition: An equal-contribution (C^{e}):

$$C_{-i,\, t-1} - \check{C}_{-i} = 0$$

Definition: An under-contribution (C^{u}):

$$C_{-i,\, t-1} - \check{C}_{-i} < 0$$

But agent i cannot assess whether agent –i over- or under-contributed with certitude because the true income of agent –i is uncertain and one cannot estimate it directly from observing actual contribution because actual contribution is partially determined by the shock.

To express the difficulty differently, let agent i starts with the assumption that agent –i has median income Y. So,

$$\check{C}_{-i} = f(\alpha Y)$$

That is, let us assume that the benchmark income of agent –i is the median income. So, any deviation from the known expected contribution can be directly attributed to the shock.

In fact, though, there is a 50% chance that the true income of agent –i is higher than Y, and 50% chance that it is lower than Y. When we take into consideration the shock, agent i actually faces four possible states of the world with 25% probability each.

	s_H with 50% prior	s_L with 50% prior
$\mathbf{y_{-i} \geq Y}$ with 50% prior	C^o	C^o & C^u
$\mathbf{y_{-i} \leq Y}$ with 50% prior	C^o & C^u	C^u

Table 1: Four States of the World

Let us say that agent i observes that the other over-contributed (C^o) in light of the assumption that the other has the median income. But agent i cannot rule that it is definitely the result of having a true income that is higher than the median ($\mathbf{y_{-i} \geq Y}$). It is possible that the true income is lower than the median ($\mathbf{y_{-i} \leq Y}$)—but the agent was blessed with good luck (s_H). Likewise, in the case of under-contribution, agent i cannot rule that it is definitely the result of having a true income that is lower than the median ($\mathbf{y_{-i} \leq Y}$). It is possible that the true income is higher than the median ($\mathbf{y_{-i} \geq Y}$)—but the agent was inflicted with bad luck (s_L).

Put differently, agent i cannot derive the conclusive inference about true income of the other by observing actual contributions. At best, agent i can use Bayes' rule to form posterior beliefs about income, i.e., probabilistic beliefs, conditioned on observed over-contributions (C^o):

$$P(y_{-i} \geq Y \mid C^o) = \frac{P(C^o \mid y_{-i} \geq Y)\, P(y_{-i} \geq Y)}{P(C^o \mid y_{-i} \geq Y)\, P(y_{-i} \geq Y) + P(C^o \mid y_{-i} \leq Y)\, P(y_{-i} \leq Y)}$$

and form posterior beliefs about income conditioned on under-contributions (C^u),

$$P(y_{-i} \leq Y \mid C^u) = \frac{P(C^u \mid y_{-i} \leq Y)\, P(y_{-i} \leq Y)}{P(C^u \mid y_{-i} \leq Y)\, P(y_{-i} \leq Y) + P(C^u \mid y_{-i} \geq Y)\, P(y_{-i} \geq Y)}$$

If income is symmetrically distributed, average income ($\tilde{Y}$) equals median income (Y). However, given the positively skewed income distribution, following assumption #9,

$$\tilde{Y} > Y$$

Given the skewness, the posteriors

$$P(y_{-i} \geq Y \mid C^o) > P(y_{-i} \leq Y \mid C^u)$$

That is, despite the fact that shocks are symmetrical, there is a greater chance to attribute over-contribution to higher income than to attribute under-contribution to lower income. This is the case because the income values to the right of the median can deviate from the median to a greater extent to which income values to the left of the median can deviate from the median. This is the consequence of the fact that the mean income is higher than the median income. So, there is a greater chance with someone with an outlier high income to offset a low shock and contribute above the expected (the median) than the case of someone with a low income to offset a high shock and contribute below the expected (the median).

Let us call the last inequality the "necessary condition," because it is necessary (but insufficient) to explain the decay of cooperation:

The necessary condition (asymmetry) of distorted belief formation:

$$P(y_{-i} \geq Y \mid C^o) > P(y_{-i} \leq Y \mid C^u)^2$$

[2] The necessary condition entails, with simplification, the following inequality expressed in terms of likelihoods of the evidence C^o and C^u:

To express the asymmetry differently, if one assesses the over-contribution of agent –i to a public good, one tends to explain it *less* in term of high shock—in comparison to the case when one assesses the under-contribution of agent –i to a public good. That is, one tends to explain more the under-contribution in relation to a low shock than explain the over-contribution in relation to a high shock.

Such asymmetry in judgment does not, by itself, engender the distorted belief formation. We need assumption #10, which states that the positive skewness of the tacit hypothesis diverges upward from true income skewness. That is, if we take the variables in ***bold and italics*** as true values and variables in regular font as the values of the tacit hypothesis held by all agents,

$$S_k > \boldsymbol{S_k}$$

$$\text{where } S_k = \frac{3(\tilde{Y} - Y)}{\sqrt{v}}$$

$$\boldsymbol{\textit{where } S_k = \frac{3(\tilde{Y} - Y)}{\sqrt{v}}}$$

$$\tilde{Y} - Y > \boldsymbol{\tilde{Y} - Y} \text{ …….. if we assume } \sqrt{v} = \sqrt{\boldsymbol{v}}$$

where S_k is Pearson's measure of skewness, $\tilde{Y}$ average income, and $\sqrt{v}$ standard deviation.

The skewness difference entails

$$P(y_{-i} \geq Y \mid C^o) - P(y_{-i} \leq Y \mid C^u) > \boldsymbol{P(y_{-i} \geq Y \mid C^o) - P(y_{-i} \leq Y \mid C^u)}$$

where $\boldsymbol{P(y_{-i} \geq Y \mid C^o)}$ and $\boldsymbol{P(y_{-i} \leq Y \mid C^u)}$ are, again, the true posteriors. That is, the difference between the posteriors of the tacit hypothesis is greater than true difference the more is the difference between the tacit hypothesis and true skewness. To recall, as discussed

$$\frac{P(C^o \mid y_{-i} \leq Y)}{P(C^o \mid y_{-i} \geq Y)} < \frac{P(C^u \mid y_{-i} \geq Y)}{P(C^u \mid y_{-i} \leq Y)}$$

earlier, the evidence does not force agents, especially with few cases of evidence, to adjust the strong tacit hypothesis concerning income distribution.

Assumption #10 can be summed up:

The sufficient condition (upward divergence of tacit hypothesis)
of distorted belief formation:

$$P(y_{-i} \geq Y \mid C^{o}) - P(y_{-i} \leq Y \mid C^{u}) > \boldsymbol{P(y_{-i} \geq Y \mid C^{o}) - P(y_{-i} \leq Y \mid C^{u})}$$

We can make two remarks: First, the right-hand side can equal zero, that is

$$\boldsymbol{P(y_{-i} \geq Y \mid C^{o}) = P(y_{-i} \leq Y \mid C^{u})}$$

which means the true distribution would be symmetrical, without violating the sufficient condition. Second, the sufficient condition entails the necessary condition, but not *vice versa*. That is, the fact that the tacit hypothesis posteriors diverge from the true ones (the sufficient condition) necessarily entails that the two posteriors are not equal (the necessary condition).

5. Judgments of Fairness: The Sufficient Condition

The sufficient condition—viz., agents believe that income is skewed more than is the case—gives rise to wrong judgments of the fairness of the contribution of others. The judgments can be either unfair or super-fair.

5.1 Unfairness Judgment

Let us suppose that agent i at period t witnesses that agent -i over-contributed in the previous period (t-1) under the assumption of having the median income. Let us also assume that t-1 is the first round of interaction. In light of the datum, agent i updates the posterior about the income of agent -i. Such update would be upward, and even more upward than what is warranted by the true skewness of income.

This entails that agent i has upgraded his belief about the unconditionally fair contribution of agent –i. Given this upgrade, agent i judges that agent –i has contributed fairly, and hence, agent i would reciprocate and contribute fairly in period t.

Meanwhile, agent –i would contribute, in period t (second round), according to his true income, which is on average lower than the posterior income believed by agent i. This is the direct implication of the sufficient condition: the believed posterior is higher than the true one. Note, Agent –i will react to the contribution of agent i in first period in the same manner in which agent i reacted—i.e., would contribute fairly in the second round.

So, while agent –i (or whoever is being assessed) is contributing fairly in the second round, others would assess the contribution as unfair given their (wrong) posterior of the income of agent –i. So, we can define unfair judgment of the contribution of agent –i as the outcome of the following inequality,

Definition: As judged by others, contribution of agent –i is unfair when:

$$E(C_{-i,\,t}) - \boldsymbol{\check{C}_{-i}\,(P(y_{-i} \geq Y \,\big|\, C^{o}_{t-1})} < E(C_{-i,\,t}) - \check{C}_{-i}\,(P(y_{-i} \geq Y \,\big|\, C^{o}{}_{,t-1}) = 0$$

The value on the right-hand states that agent –i, as truly the case, contributes on average in the second round according to true income. But the value on the left-hand is negative because the assessed unconditional fair contribution, as a function of the posterior, is higher than the true one.

5.2 Super-fairness Judgment

This is the mirror-image of the unfair judgment. Let us suppose that agent i at period t (second round) witnesses that agent -i under-contributed in the previous round (t-1) under the assumption that agent -i enjoys the median income. In light of the datum, agent i updates the posterior about the income of agent -i. Such update would be downward—and even more downward as a result of beliefs concerning skewness of income is greater than true skewness.

As before, this entails that agent i has upgraded his belief about the unconditionally fair contribution of agent –i. Given this upgrade, agent i judges that agent –i has contributed fairly, and hence, agent i would reciprocate and contribute fairly in period t.

Meanwhile, agent –i would contribute, in the second round, according to the *warranted* posterior, which is higher than what is believed by agent i. Note, Agent –i will react to the contribution of agent i in first period in the same manner in which agent i reacted—i.e., would contribute unconditionally fair contribution in the second round. So, the only source of discrepancy is the sufficient condition: the downward posterior is lower than the true one. So, while agent –i (or whoever is being assessed) is contributing fairly in the second round, others would assess the contribution as super-fair given their (wrong) posterior of agent –i, which is lower than the true one.

So, we can define super-fair judgment of the contribution of agent –i as the outcome of the following inequality,

Definition: As judged by others, contribution of agent –i is super-fair when:

$$E(C_{-i,\,t}\,)\ \boldsymbol{-\ \check{C}_{-i}\ (P(y_{-i} \leq Y \mid C^{u}_{,t-1})}\ >\ E(C_{-i,\,t}) - \check{C}_{-i}\ (P(y_{-i} \leq Y \mid C^{u}_{,t-1}) = 0$$

This definition is similar, but in reverse, of the previous one.

5.3 Fairness Judgment

In light of the above, fairness judgment will arise when the believed posterior matches the true posterior. Then, the assessed unconditionally fair contribution of others, as a function of the posterior, would match the true assessed value.

Definition: As judged by others, contribution of agent –i is fair when:

$$E(C_{-i,\,t}\,)\ \boldsymbol{-\ \check{C}_{-i}\ (P(y_{-i}\ \geq\ Y \mid C^{o}_{,t-1})}\ =\ E(C_{-i,\,t}) - \check{C}_{-i}\ (P(y_{-i}\ \geq\ Y \mid C^{o}_{,t-1}) = 0$$

$$\text{where } \boldsymbol{\check{C}_{-i}\ (P(y_{-i}\ \geq\ Y \mid C^{o}_{t-1})}\ =\ \check{C}_{-i}\ (P(y_{-i}\ \geq\ Y \mid C^{o}_{t-1})$$

$$E(C_{-i,\,t}\,)\ \boldsymbol{-\ \check{C}_{-i}\ (P(y_{-i} \leq Y \mid C^{u}_{,t-1})}\ =\ E(C_{-i,\,t}) - \check{C}_{-i}\ (P(y_{-i} \leq Y \mid C^{u}_{,t-1}) = 0$$

$$\text{where } \boldsymbol{\check{C}_{-i}\ (P(y_{-i} \leq Y \mid C^{u}_{t-1})}\ =\ \check{C}_{-i}\ (P(y_{-i} \leq Y \mid C^{u}_{t-1})$$

That is, in either case, when we have either over- or under-contribution, actual contribution matches the unconditionally fair contribution on average. This is possible only when the believed posterior matches the true posterior, i.e., when the sufficient condition of upward divergence of tacit belief does not hold.

6. Distorted Belief Formation: The Necessary Condition

In case that the tacit hypothesis diverges from true skewness, agents would be forming super-fairness and unfairness judgments (beliefs) concerning the contribution of others—but there would be no fairness judgments. If so, would the super-fairness judgments totally offset the unfairness judgments? If they cancel each other out perfectly, the agent under-focus should form the opinion that, on average, others are contributing fairly and, hence, he or she would make the conditionally fair contribution (i.e., act as a perfect conditional cooperator).

To recall, though, the necessary condition, income is skewed to the right and, hence,

$$P(y_{-i} \geq Y \mid C^{o}) > P(y_{-i} \leq Y \mid C^{u})$$

This entails, as illustrated in Table 2, for unfairness judgments to exceed super-fairness

	s_H with 50% prior	s_L with 50% prior
$\mathbf{y_i \geq Y}$ with 50% prior	C^o with 25% likelihood	C^o with σ likelihood C^u with (.25-σ) likelihood
$\mathbf{y_i \leq Y}$ with 50% prior	C^o with μ likelihood C^u with (.25-μ) likelihood	C^u with 25% likelihood

Table 2: Likelihoods of Over- and Under-Contribution

judgments. Table 2 is a modification of Table 1, where obviously $0<\sigma\leq25\%$ and $0<\mu\leq25\%$. Let us examine the state of the world of $y_i \geq Y$ (income higher than median) and s_L (low income shock). As Figure 1 shows, anyone with a low-end of the range Y would under-contribute, while

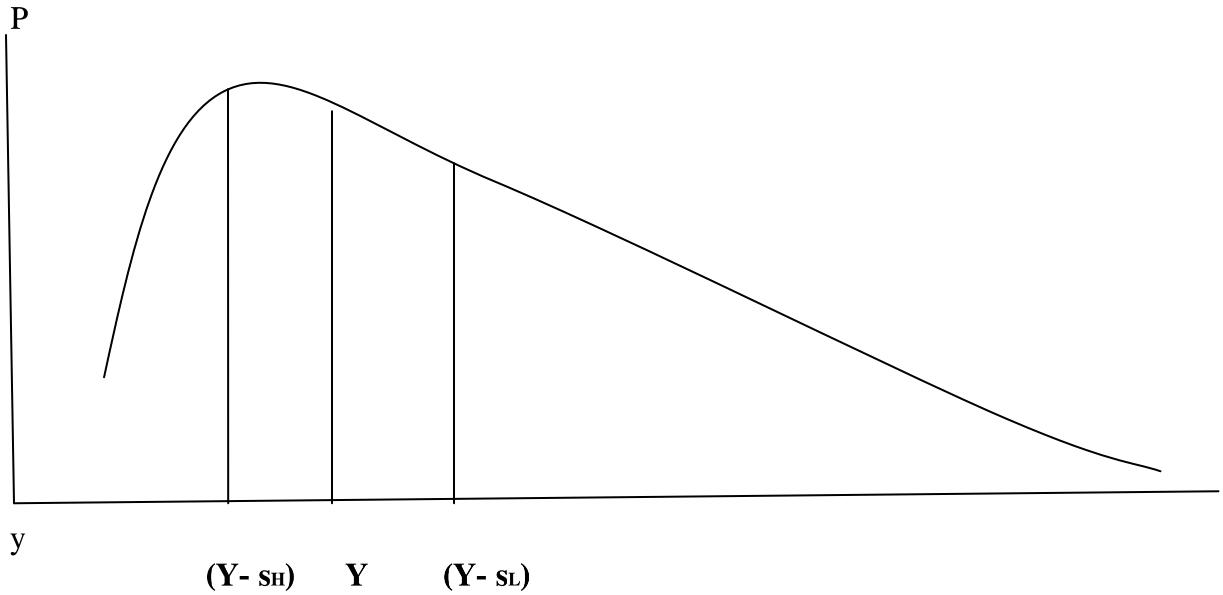

Figure 1: $s_H = - s_L$

anyone with a high-end of the range would over-contribute. Precisely,

$$Y \leq y_i \leq (Y - s_L) \rightarrow C^u$$

$$y_i > (Y - s_L) \rightarrow C^o$$

$$\text{where } s_L < 0$$

On the other hand, as shown in Figure 1, in the state of the world of $y_i \leq Y$ and s_H, anyone with a low-end of the range would under-contribute, while anyone with a high-end of the range would over-contribute. Precisely,

$$Y \geq y_i \geq (Y - s_H) \rightarrow C^o$$

$$y_i < (Y - s_H) \rightarrow C^u$$

$$\text{where } s_H > 0$$

To see which is more likely, $P(C^u)$ or $P(C^o)$, we need to keep in mind that $s_H = - s_L$. Without loss of generality, as Figure 2 shows, let us assume that s_H (and hence s_L) is as

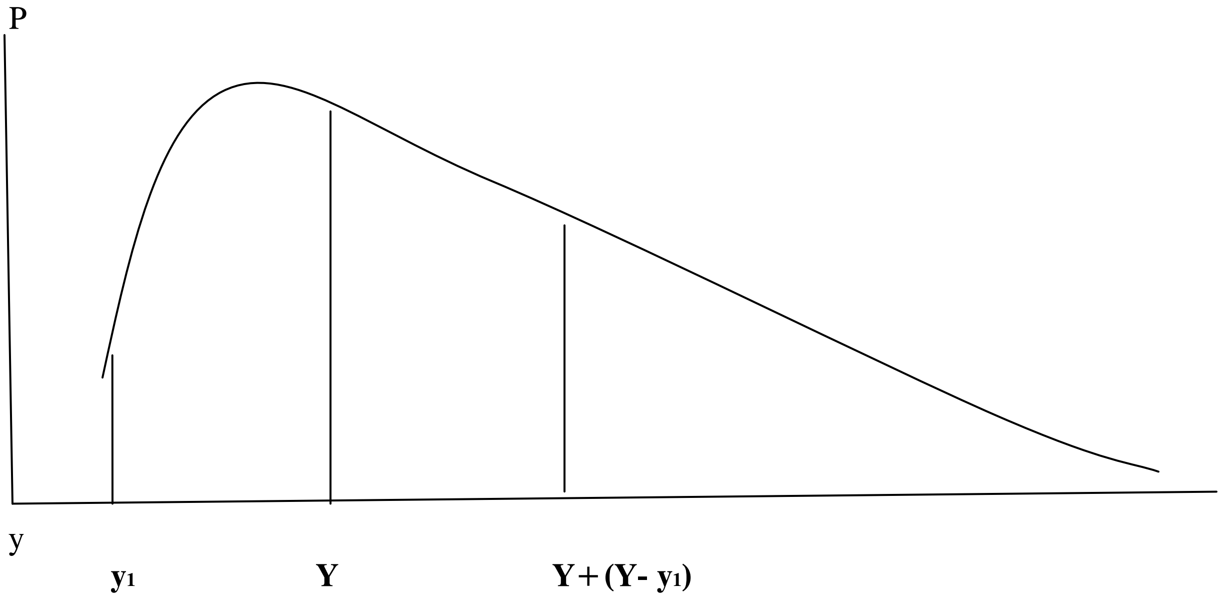

Figure 2: Shock = $(Y\text{-}\ y_1)$

large as $(Y\text{-}\ y_1)$, wher y_1 is the income of the poorest agent in the population. In the state of the world of $y_i \leq Y$ and s_H, and where $s_H = (Y\text{-}\ y_1)$:

$$P(Y \geq y_i \geq Y - (Y\text{-}\ y_1)) \rightarrow P(C^o) = 25\%$$

$$P(y_i < Y - (Y\text{-}\ y_1)) \rightarrow P(C^u) = 0$$

In contrast, in the state of the world $y_i \geq Y$ and s_L, there will be, as Figure 2 shows, individuals whose income is higher than $Y+ (Y\text{-}\ y_1)$. So, the likelihood of over- and under-contribution, where $s_L = - s_H = - (Y\text{-}\ y_1)$,

$$P(y_i > Y + (Y\text{-}\ y_1)) \rightarrow P(C^o) = \sigma$$

$$P(Y \leq y_i \leq Y + (Y\text{-}\ y_1)) \rightarrow P(C^u) = 25\% - \sigma$$

If we add up the frequencies of $P(C^u)$ and $P(C^o)$ in both states of the world:

$$P(C^o) + P(C^o) = 25\% + \sigma$$

$$P(C^u) + P(C^u) = 25\% - \sigma$$

That is, the frequency of judgments of over-contribution exceeds the frequency of judgments of under-contribution in both states of the world. Even when we generalize with shocks that are smaller than $(Y- y_1)$, we would reach same result.

In the two other possible states of the world, in the first and fourth quadrants of Table 2, $P(C^u)$ and $P(C^o)$ fully offset each other. So, in total,

$$P(C^o) - P(C^u) = 2\ \sigma$$

This asymmetry of over-contribution is simply the outcome of the necessary condition: positively skewed distribution. It is not the sufficient condition for the decay of cooperation. If only the necessary holds, i.e., if the believed skewed distribution matches the true one, agents would form the correct beliefs, i.e., would reach the judgment that the contributions of others are fair and, hence, would never retaliate by lowering their own contribution below the unconditionally fair contribution level. That is, even if the frequency of over-contribution is greater than the frequency of under-contribution, it is insufficient to account for the decay of cooperation. We need the sufficient condition: The believed skewness is greater than the true one. As discussed above, this leads to the incidence of super-fair and unfair judgments. But now, given the necessary condition (i.e., $P(C^o) - P(C^u) = 2\ \sigma$), the unfair judgments exceed the fair judgments.

Consequently, each agent come to believe that others, on average, are not contributing fairly--which is not the case in the model. But given the assumption that contribution is conditional on the belief of what others are contributing, agents wrongly retaliate by reducing correspondingly their contribution. This sets in motion the decay of cooperation.

The decay may take a form different from being disappointed with the contribution of the early contributors. It can also take the form of giving more leeway, tolerance, sympathy with late contributors under the justification that they are under stress and need a break. Such sympathy would fit the proverb that "the squeakiest wheel gets most of the oil"—i.e., people who complain the most or are late in their contributions get (wrongly) greater relief.

7. Significance and Payoffs

The proposed hypothesis, which relies exclusively on the role of beliefs rather than odd preferences, has at least four payoffs.

7.1 Experimental Laboratory Setting

The proposed explanation of the decay of cooperation might explain why laboratory experiments, with small numbers of participants, lead to the decay of cooperation: The tacit hypothesis may reflect true income skewness in the larger population, while the income distribution of participants in the laboratory is close to a normal given the small number of the sample and the sample is drawn from university student population. Such a population, in any university, involves self-selection, where each university attracts students with homogeneous income levels.

7.2 Restart Effect

The proposed agenda of starting with beliefs, rather than preferences, can easily explain the restart effect puzzle. If we start with the preference of imperfect conditional cooperation, the one would “free ride” more in succeeding sessions as a result of learning—given that the motive is already imperfect, i.e., one wants to cheat a little bit. But if start with beliefs, and assume that one is fair (perfect conditional cooperator), one is tacitly aware that the collapse is not a result of suboptimal preferences (“bad” intention) on one’s part or on the part of others, but rather the outcome of mis-coordination. Thus, one would be determined, in a fresh session, to start again with the same level of commitment to the optimal contribution, the fair contribution. This interpretation actually confirms the notion of some philosophers [e.g., Gauthier 1986] that the cooperation strategy is social dilemma is not only the Pareto optimum but also the rational (Nash equilibrium) outcome.[3] So, despite previous experiences, participants start with high contributions as an attempt to nullify or correct the distorted belief formation.

7.3 Institutions

The proposed hypothesis of distorted belief formation can shed light on many institutions erected throughout history to nullify or correct such formation. There are at least three kinds of institutions worth mentioning:

7.3.1 Societies have opted for “third parties” to decide on the proper punishment, as the case of

[3] In fact, for Gauthier [1986], the Pareto optimum of the prisoners’ dilemma or, in general, public good games is the *uniquely* rational strategy [see Gauthier & Sugden, 1993; Khalil, 1997]. That is, rational agents would not defect. However, for Bacharach [2006; see Gold & Sugden, 2007; Sugden, 2000, 2003], the Pareto optimum is only rational if one is motivated by one presentation or frame, viz., the collective “we-intention” that arises when one acts as a member of a team. But it is not rational (the view arising from subgame perfect Nash equilibrium) if one is motivated by another presentation or frame, viz., the self-interest “me-intention.”

murder in the Arabian Desert recounted at the outset of this proposal. If agents are left on their own, and they are aware that any act of contribution or any act of remedy an injury can be misinterpreted by the others, they would hold back even on simple acts of cooperation. For instance, if agents are left on their own without third parties, they would rather act "stubborn" and avoid any admission of guilt. They would be suspicious that the injured party would find them "weak," i.e., find that the "early admission of guilt" as a signal of true indebtedness. Then, the injured part would feel it is only "fair" to ask for greater compensation.[4]

7.3.2 Mediators, such as real estate agents, act as institutions that mitigate the distorted belief formation problem. The mediators can be seen as judges that try to avoid the distorted belief formation process. If traders are left on their own to bargain they may not reach an agreement even when there are many possible agreements that are Pareto-improving. The mediators, as buffers in bargaining, can find such possible Pareto-improving agreements without making one party exposed to pressures by the other for more concessions.

7.3.3 Seed money and leadership-initiatives are a third kind of institution aimed at reversing the distorted belief formation. When an agent nominate him or herself to put seed money, he

[4] This may explain the situation where it is known that ultimatum in international relations is usually the standard excuse to start a war. When one party issues an ultimatum, it is almost certain that the other party, even if it is much weaker, would never concede publicly because such concession would mean "early admission of guilt"—which would lead to further pressures. For example, in the build up to the first Gulf war in Jan-March 1991, the UN's Security Council issued an ultimatum to Iraq to withdraw from Kuwait. Iraq rejected the ultimatum, because the UN is not actually a third party that can mediate. It has no army to defend Iraq to stop further pressures and demands put up by the US and its allies. This explains the benefit of using credible referees or judges, whose ruling can be obeyed by the perpetrator (Iraq) without inviting further punishment by the injured party (Kuwait and the US).

or she is in effect declaring that the action is not to be interpreted as a “fair contribution.” If he or she makes the contribution as usual, others would most likely expect him or her to contribute even more judged by the unconditionally fair contribution that the contribution is motivated by fairness. And this would lead to decay of cooperation as discussed above. However, if the contribution is set up against the unconditionally fair contribution that it is “seed money,” then others would not judge it according to the common norm of fairness. In this manner, it rather acts as a catalyst to sustain belief formation contrary to its natural course of decay.

7.4 Beyond the Free-Riding Problem

The theory of public goods is dominated by the free-riding problem. Even the idea of the odd preference, viz., imperfect conditional cooperation, is laced by the free-riding problem. Likewise, the dominant view of organization economics is the agent-principal framework that stresses the role of opportunism and the best design of governance needed to combat opportunism. The proposed hypothesis offers an alternative explanation that stresses the role of mis-coordination. If so, free-riding might not be the major reason for the under-provision of public goods. At least analytically, as this paper shows, the under-provision of public goods and the decay of cooperation may rather stem from the coordination problem, namely, the tendency of agents to over-expect the contribution of others. If so, one payoff of this research is that the solution of the under-provision of public goods resides, at least partially, in greater information and coordination, rather than greater monitoring.

8. Testing

It is possible to test experimentally the income skewness hypothesis. We have two theories behind the collapse of cooperation: distorted belief formation as a result of skewed income or imperfect conditional cooperation. The latter does not have to be wrong in order for the former to be correct. So, the experiment only needs to show that there is some support for the proposed hypothesis.

While participants have a general idea of the population's income skewness, they do not realize that the skewness is smaller in the small sample in the laboratory. In fact, the income distribution is equal in the laboratory setting, while participants are using, tacitly, the tacit hypothesis of distribution of the population at large.

The experiment can be broken into two experiments:

1. The baseline replicates the standard set up of the public good game. Then, the treatment consists of making participants fully aware of the equal distribution of endowment among the participants. The experimenter then quizzes them about their tacit hypothesis concerning the income distribution of the population. Then the experimenter asks them about the extent of the gap between the endowment distribution in the lab and the true population income distribution. In this manner, participants become fully aware of their tacit hypothesis of the income distribution. Thus, it is predicted that participants would ignore the tacit hypothesis and, when they play the public

good game again, cooperation would be significantly better than cooperation in the baseline, even if it eventually collapses.

2. The baseline replicates the standard set up of the public good game. Then, the experimenter increases by one- and, in other treatments, by many-fold the endowment of one or a few participants, while holding the endowment of others constant. In one treatment, participants are made aware of the changing income distribution, but unaware which particular players are made richer. In another treatment, participants who receive no increases are not informed of the income increases of the others, while the others who experience income rise are told that the increase is only limited to them. In both treatments, income skewness in the lab is manipulated to approach the supposed tacit hypothesis about income skewness. The prediction is that we should witness, against the baseline, an increase in cooperation in the case when some participants are kept unaware of the endowment manipulation. As to the case when all are aware, the direction of cooperation would be unpredictable, given the fact that the level of expectation (fairness) might rise as a result of being aware of the distribution.

9. Conclusion

This proposal offers an alternative explanation of the decline of cooperation. It declines not because of imperfect preferences concerning conditional cooperation, but rather because of distorted belief formation.

The proposal argues that there are two conditions behind the distorted belief formation:

1. **Necessary Condition:** Income distribution is positively skewed;
2. **Sufficient Condition:** The tacit hypothesis held by participants in public good games concerning the income distribution is more positively skewed than the true income distribution of the participants.

The proposal shows how a wrong tacit hypothesis, given the confines of a situation that it cannot be corrected, can lead to mis-coordination, misunderstanding, and hence the decay of cooperation. The proposal shows the important payoffs of the idea that distorted belief formation is behind the decay of cooperation. It shows how societies erect institutions (such as third parties, attorneys, and leadership) in order to avoid the distorted belief formation. It is crucial for a society, where cooperation is crucial, to avoid the tendency for cooperation to decay. Cooperation is the basis of division of labor, which is the source of wealth of nations as stressed by Adam Smith. Designers of governance structures should pay greater attention to sustaining clear communication and coordination, rather than the current emphasis on monitoring and policing, in order to sustain cooperation and ensure the further development of division of labor.

References

Alchian, Armen A. and Harold Demsetz. "Production, Information Costs, and Economic Organization." *American Economic Review*, December 1972, 62:5, pp. 777-795.

Andreoni, James. “Cooperation in Public-Goods Experiments: Kindness or Confusion?” *American Economic Review*, September 1995, 85:4, pp. 891-904.

Bacharach, M. *Beyond Individual Choice: Teams and Frames in Game Theory,* edited by Natalie Gold and Robert Sugden. Princeton: Princeton University Press, 2006.

Binmore, Kenneth. “Economic Man—or Straw Man?” *Behavioral and Brain Sciences*, December 2006, 28:6, pp. 817-818.

Botelho, Anabela, Glenn W. Harrison, Lígia M. Costa Pinto, and Elisabet E. Rutström. “Testing Static Game Theory with Dynamic Experiments: A Case Study of Public Goods.” *Games and Economic Behavior*, 2009, 67, pp. 253–65.

Chwe, Michael Suk-Young. *Rational Ritual: Culture, Coordination, and Common Knowledge.* Princeton, NJ: Princeton University Press, 2001.

Chaudhuri, Ananish, Andrew Schotter, and Barry Sopher. “Talking Ourselves to Efficiency: Coordination in Inter-Generaltional Minimum Effort Games with Private, Almost Common and Common Knowledge of Advice. ” A working paper, Department of Economics, University of Auckland, 2007.

Fischbacher, Urs, Simon Gächter and Ernst Fehr. “Are People Conditionally Cooperative? Evidence from a Public Goods Experiment.” *Economics Letters*, 2001, 71, pp. 397-404.

Fischbacher, Urs and Simon Gächter. “Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Good Experiments.” *American Economic Review*, March 2010, 100:1, pp. 541-556.

Frank, Robert H. *Passions Within Reason: The Strategic Role of the Emotions*. New York: W.W. Norton, 1988.

Gächter, Simon. “Conditional Cooperation: Behavioral Regularities from the Lab and the Field and Their Policy Implications.” CeDEx Discussion Paper No. 2006-03, April 2006.

Gauthier, David. *Morals by Agreement*. Oxford: Oxford University Press, 1986.

________ and Robert Sugden, eds. *Rationality, Justice and the Social Contract: Themes from ‘Morals by Agreement’*. London: Harvester Wheatsheaf, 1993.

Gold, Natalie and Robert Sugden. “Collective Intentions and Team Agency.” *Journal of Philosophy*, 2007, 104, pp. 109-137.

Hirshleifer, Jack. “On the Emotions as Guarantors of Threats and Promises.” In John Dupré (ed.) *The Latest on the Best: Essays on Evolution and Optimality*. Cambridge, MA: MIT Press, 1987, pp. 307-326.

Houser, Daniel and Robert Kurzban. “Revisiting Kindness and Confusion in Public Goods Experiments.” *American Economic Review*, September 2002, 92:4, pp. 1062-1069.

Isaac, Mark R., Kenneth McCue, Charles R. Plott. “Public Goods Provision in an Experimental Environment.” *Journal of Public Economics*, 1985, 26, pp. 51-74.

Khalil, Elias L. “The Red Queen Paradox: A Proper Name for a Popular Game.” *Journal of Institutional and Theoretical Economics*, June 1997, 153:2, pp. 411-415.

________. “Tastes.” *International Encyclopaedia of the Social Sciences*, 2nd ed. Edited by William A. Darity, Jr., Vol. 8. Detroit: Macmillan Reference USA, 2008, pp. 266-270.

________. “Self-Deceit and Self-Serving Bias: Adam Smith on ‘General Rules`.” *Journal of Institutional Economics*, August 2009, 5:2, pp. 251-258.

Kreps, David M. and Robert Wilson. "Reputation and imperfect information." *Journal of Economic Theory*, August 1982, 27:2, pp. 253-279.

Kreps, David M., Paul Milgrom, John Roberts and Robert Wilson. “Rational Cooperation in the Finitely Repeated Prisoners’ Dilemma.” *Journal of Economic Theory*, 1982, 27, pp. 245-252.

Kurzban, Robert, and Daniel Houser. “An Experimental Investigation of Cooperative Types in Human Groups: A Complement to Evolutionary Theory and Simulations." *Proceedings of the National Academy of Science USA*, 2005, 102:5, pp. 1803-1807.

Lawrence, T.E. *Seven Pillars of Wisdom*, intro. by Angus Calder. Wordsworth Classics of World Literature, 1997.

Olson, Mancur. *The Logic of Collective Action.* Cambridge, MA: Harvard University Press, 1965.

Stigler, George J. and Gary S. Becker. "*DE Gustibus Non Est Disputandum*." *American Economic Review*, March 1977, 67:1, pp. 76-90.

Sugden, Robert. “Team Preferences.” *Economics and Philosophy*, 2000, 16, pp. 175-204.

________. “The Logic of Team Reasoning.” *Philosophical Explorations*, 2003, 6, pp. 165-181.

Young, H. Peyton. "The Economics of Convention." *Journal of Economic Perspectives*, Spring 1996, 10:2, pp. 105-22.

________. "The Evolution of Conventions.” *Econometrica,* January 1993, 61:1, pp 57-84.

Appendix:
How Does Imperfect Conditional Cooperation Work?

Fischbacher and Gächter [2010] find that while people are ready to contribute to a public good in finite games, they contribute, first, conditionally on the contribution of others and, second, they tend to cheat a bit. Thus, they call them *imperfect* conditional cooperators. If the predicted contribution, e.g., is $10 per participant on average, the average participant would contribute something below $10—not zero as predicted by subgame perfect Nash equilibrium and not $10 as predicted by prosocial preferences of perfect conditional cooperaiton. That is, most participants are neither *perfect* free-riders nor *perfect* conditional cooperators, denoted as “pCC” by Fischbacher/Gächter. Participants are rather quasi free-riders or, what Fischbacher/Gächter call, *actually* imperfect conditional cooperators, denoted as “aCC.”

To test their conjecture, they set up an experiment where they elicit the “preferences” of aCC. Then they let the preferences interact with the “belief” of what is the predicted contribution of others in each round of the game. They want to find out what is the primary motor behind the decay of cooperation—is it the elicited preference or is it the ever-changing belief?

For this purpose, Fischbacher/Gächter set up a two-stage experiment, one to capture the preference to contribute and the other to capture the belief. Concerning capturing the preference, they call it “P-experiment,” they elicit each participant’s preference to contribute. They ask each participant what amount of money he or she would be ready to contribute for each level of contribution. So, they generate a schedule for each participant. While they find many perfect free-riders, and a few unconditional cooperators and perfect conditional cooperators (pCC), the majority are conditional cooperators with a hint of cheating, what they call “*actual* conditional cooperators” (aCC).

To capture actual *contribution* in light of the belief, called “C-experiment,” they design ten one-shot standard public good games with random matching, and repeated the set up in six sessions (three sessions start with P-experiment followed by the C-experiment, and the other three in reverse). Each shot is composed of 4 participants that are continuously re-matched to reduce the incentive to cooperate in order to buttress high belief and hence high contributions by others, for which there is some evidence [Botelho *et al.*, 2009]. They had 24 participants in each of the five sessions, while had 20 participants in the sixth session. So, they had a total of 140 participants. As Fischbacher/Gächter [2010, p. 544, n.6] report: “The likelihood in period 1 that a player would meet another player once again during the remaining nine periods was 72 percent. The likelihood that the *same* group of four players would meet was 2.58 percent. Since the experiment was conducted anonymously, however, subjects were unable to recognize whether they were matched with a particular player in the past.”

In each shot, before any contribution is made, the participants express their belief about their estimate of the average contribution of the three other players. And participants are rewarded for providing close estimates of the average contribution. Further, there was no sequence order effect: Whether participants started with the P-experiment followed by the C-experiment, or *vice versa,* did not influence their contribution.

To understand the interaction of beliefs and preferences, they define two kinds of belief formations:

Naïve belief formation: $\text{belief}_t = \text{Contribution}_{t-1}$

But they find that participants were involved in actual (non-naïve) belief formation:

Actual belief formation: belief$_t$ = average (Belief$_{t-1}$ + Contribution$_{t-1}$)

But how is the contribution related to the belief in each round? As Figure 2 clarifies,

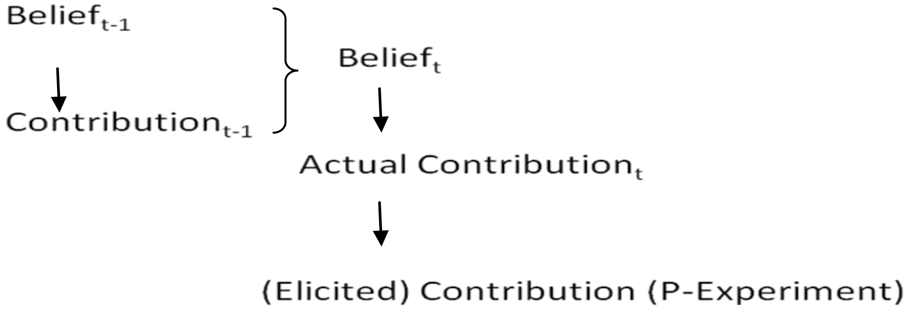

Figure 2: Belief Formation

the contribution in C-experiment (similar to P-experiment) is lower than the corresponding beliefs of same round. Given that the belief of the succeeding round is somewhere between the belief and contribution of the previous round, beliefs tend to decline over rounds.

Further, as Figure 2 shows, the actual contribution in each period, given the belief, is lower than the contribution specified in the elicited preferences (P-experiment). That is, the belief of previous round somewhat makes the actual contribution higher than what is expressed in the elicited preferences (P-experiment).

Given the tendency of beliefs and contributions to decline, cooperation is destined to collapse. But what is the main factor behind the collapse? Is it declining beliefs or is it the imperfect preferences (aCC)?

Fischbacher/Gächter address this question by performing a simulation using the parameters of the experiment. The simulation allows them to disentangle the declining belief from the preference. The non-naïve belief simply slows down the decay of the contributions. But the role of the belief, whether formed naively or non-naively is totally nullified if participants are pCC. That is, if participants are ready to contribute what others contribute on average, the belief (whether non-naïvely or naïvely) formed has no effect on the decay of cooperation. (They also find no effect for the issue of heterogeneity of prosocial preferences).

That is, Fischbacher/Gächter find that the motor behind the decline of cooperation is not the formation of beliefs but rather the preference “aCC,” i.e., agents are imperfect conditional cooperators.

But are we sure that participants are imperfect conditional cooperators (aCC), as found by the P-experiment? When participants were asked about their reaction to specific schedule of contributions, they were not asked as to why they chose amounts lower than perfect conditional cooperation. If asked, they may reveal that they have an expected contribution, something that corresponds to what is fair, and hence find the amount presented at the schedule to be lower than the fair amount. So, when participants were told about a specific schedule in the P-experiment, they reciprocated by a lower amount probably because they were judging the actual schedule against the fair contribution.