

Social Trust, Cooperation, and Human Capital

Fali Huang*

Department of Economics
The University of Pennsylvania

March 3, 2003

Abstract

The importance of social trust on economic growth has been suggested by many empirical works. This paper formalizes the concept of social trust and studies its formation process in a game theoretic setting. It provides plausible explanations for a wide range of empirical and experimental findings. The main results of the paper are as follows. Social trust in a game is determined by the distribution of players' cooperative tendency and specific game features. It induces cooperation even in one-period prisoner's dilemmas and strictly increases the total outputs. Cooperative tendency in essence is a component of human capital distinct from cognitive ability. People may choose to invest in appropriate cooperative tendencies to maximize their life-time utility. The investment, however, is typically not efficient because its social returns are always strictly larger than individual returns. Multiple equilibria exist so that social trust levels are history dependent. However, the stable equilibrium with non-extreme social trust level is unique when it exists. The model clarifies the roles of various forces in affecting social trust. In particular, efficient disciplinary institutions such as legal system increase social trust, but may crowd out the average cooperative tendency of players. And investing in cooperative tendency may actually induce people to choose higher cognitive abilities than otherwise. (*JEL* Z13, J24)

1 Introduction

As Arrow (1972, p.357) has observed, "Virtually every commercial transaction has within itself an element of trust, certainly any transaction conducted over a period of time. It can

*I am grateful to George Mailath, Andrew Postlewaite, and Rafael Rob for their support. I also thank Hanming Fang, Mokoto Hanazono, Volker Nocke, Dan Silverman, and Huanxing Yang for comments and suggestions. All remaining errors are mine.

be plausibly argued that much of the economic backwardness in the world can be explained by the lack of mutual confidence.” Arrow’s insight about the importance of social trust in economic life has been confirmed by many recent empirical works. For example, the average trusting level in a society, measured by *TRUST* based on the World Value Surveys, has large positive effects on performance of various organizations (La Porta et al. 1997). It is also significantly associated with economic growth (Knack and Keefer 1997).

Social trust as an important social phenomenon has been extensively studied by social scientists (see for example Cook 2001). In recent years it attracts attention from economists as one of the most important forms of social capital. Parallel to physical and human capital, the term ‘social capital’ is created to represent the cooperative infrastructure of a society (Coleman 1988, Putnam 1993, 1995). It often refers to features of social organization such as networks, norms, and social trust that facilitate coordination and cooperation for mutual benefit (Putnam 1995).

We initially aimed at studying the formation of social capital in a society. But more thoughts on it suggest that social capital is an umbrella concept whose many manifestations substantially differ from and interact with each other. They share the same name social capital because all of them belong to the not yet well-appreciated social forces that constitute the cooperative infrastructure of our society, which does not mean that they are qualitatively or analytically homogenous. Among the various forms, we find that social trust best captures the essence of social capital. And rigorous studies of social trust may be the key to help us thoroughly understand social capital.

This paper formalizes the concept of social trust and studies its formation in a society using a game theoretic setting. A key insight is that one-period prisoner’s dilemma is the ideal context to formalize social trust. Note that the concept of *trust* is vacuous without the discrepancy between social and individual returns, since otherwise rational people can always be ‘trusted’ to choose their optimal actions. And *social* trust is typically referred to trust among strangers, rather than acquaintances that are involved in repeated interactions.

The paper shows that all trust related concepts can be built from a single analytical

element: cooperative tendency. People in general have heterogenous cooperative tendencies (tastes for cooperation). Fix a one-period prisoner's dilemma. Players with sufficiently high (low) cooperative tendencies will always cooperate (defect), while those in the middle behave reciprocally. A person's *trustworthiness* in this game, defined as the probability that he will cooperate in it, is characterized by his cooperative tendency. How much *trust* people put in a specific person in this game is equal to his trustworthiness. The *social trust* in a group is the expected trustworthiness of a typical member, determined by the distribution of cooperative tendency in the group.

Across games, a person with a certain cooperative tendency may have different levels of trustworthiness, while those with higher cooperative tendencies are in general more trustworthy. Social trust also varies across games and decreases with defecting benefits.

The analysis shows that cooperation exists in one-period prisoner's dilemmas if and only if there is social trust among players. The total outputs strictly increase with the amount of social trust. Since discrepancies between social and individual returns captured by prisoner's dilemmas are quite common in our society, social trust can enhance organization performance and overall economic growth.

A simple social trust formation model is developed where cooperative tendencies are endogenized. Parents, taking as given the expected social trust in the society, may help their children develop appropriate cooperative tendencies in order to maximize their lifetime utility. Social returns to cooperative tendency investment are always strictly larger than individual returns, implying that the equilibrium social trust is very likely to be inefficient.

In the benchmark case where individual returns of cooperative tendency are moderate and quite similar among players, a negligible difference in initial beliefs may lead the economy to either 'no trust' or 'full trust' stable equilibrium. In contrast, when the returns are diverse, there is only one stable equilibrium. Multiple equilibrium is also common in other cases, while the stable equilibrium with non-extreme social trust level is unique when it exists.

The model suggests that social trust in stable equilibrium can be improved by increas-

ing the efficiency of information structure, formal and informal disciplinary institutions, and cooperative tendency development process. Dense social networks and mass media, by facilitating information flowing, can increase the returns to cooperative tendency. Families, communities, and schools are the main forces that affect the investing cost of cooperative tendency in a society. Examples of disciplinary institutions are legal systems, monitoring schemes, social networks, and social norms. When the efficiency of these institutions increases, social trust goes up, but the average cooperative tendency of players goes down.

In an extension to the basic model, the investment in cooperative tendency, as a part of a child's social development, is optimally balanced against investment in child cognitive development. These two kinds of investment compete with each other for limited resources. On the other hand, however, both may improve individuals' life-time incomes. In this sense cooperative tendency is another component of human capital besides cognitive ability. We find that, when the cost is low enough, investing in cooperative tendency may induce people to choose higher cognitive abilities.

The paper contributes to the emerging social capital literature in several aspects. First, it formalizes the concept of social trust and subjects it to rigorous analysis. It also studies how social trust is formed in the society. This formalization may be helpful in the building of a unified conceptual framework in which rigorous analysis of all forms of social capital can be carried out. For example, the paper shows that social network and norms interact with current social trust in promoting cooperation, and they affect the formation of future social trust. On the other hand, as long as discrepancies between social and individual returns are involved, social trust may also play a crucial role in the creation and maintenance of these social capital forms.

Next, the paper accounts for many empirical and experimental findings about social trust. For example, 1) social trust improves economic performance at various levels, 2) the discrepancies between different measures of social trust are quite common; 3) the social trust levels can be quite different in otherwise identical groups; 4) social trust is positively associated with mass media, formal institutions, and dense social networks; 5) strangers

rationally cooperate in public goods experiments, and 6) they behave differently across games, while the behaviors of the same person share some stable characteristics.¹

Finally, the paper also yields a number of testable implications. For example, 1) there exist strong positive externalities in cooperative tendency investment; 2) cooperative tendency and cognitive ability may be complements or substitutes for different people; 3) important links exist between people's cooperative tendencies, education system, and social trust; 4) careful choices of subjects and defecting benefits in experiments are crucial to get unbiased estimates of social trust in a larger group.

An earlier attempt to study social capital formation is made by Glaeser, Laibson, and Sacerdote (2000). The current paper differs from theirs in a couple of important aspects. First, they bundle different forms of social capital together as a homogenous subject, while we focus on social trust and clarifies its relationship with other forms. Second, they model individual players' investment decisions as isolated optimization problems, while we study social trust formation as an equilibrium result to capture the important externality among players.

Another related work is by Rob and Zemsky (2002). They study the effects of incentive structures in affecting social trust at the firm level, assuming that employees' cooperative tendencies change mechanically. The current paper complements their work in that we focus on individual optimal choices in cooperative tendency and study the formation of social trust in a society.

The paper suggests a link between social trust and reputation literature. It shows that the existence of social trust is a necessary condition for reputation effects to exist among utility-maximizing players. In other words, the role of repeated interactions in promoting cooperation is not generating new social trust, but increasing the effects of *existing* social trust.² Furthermore, the formulation of cooperative tendency naturally *derives* some crucial types of players in achieving cooperation that are typically *assumed* exogenously given (see,

¹Detailed evidence is available from Knack and Keefer (1997), La Porta et al. (1997),

²Repeated interactions may generate emotional ties or altruistic feelings that also promote cooperation. But this is another story other than *social* trust.

e.g. Kreps et al. 1982, Tirole 1996, Fehr and Gächter 2000).

This paper is also closely related to the human capital literature, where several recent works provide evidence that non-cognitive skills, including incentive-enhancing preferences, are important in determining individual earnings (Heckman 2000, Bowles et al. 2001). The current paper is distinct from these works in that we adopt an aggregate approach and thus detect the strong positive externality in the investment of cooperative tendency. It suggests that investment inefficiency may be more severe for non-cognitive skills than for cognitive ones.

The paper is organized as follows. In the next section cooperative tendency and social trust are formally defined, their effects on individual and aggregate outputs are analyzed. A simple social trust model is developed in section three, where players' cooperative tendencies are chosen (by their parents) to maximize their life time utility. The comparative statics and their empirical implications in improving social trust are also discussed. The final section presents conclusions.

2 The Formalization of Social Trust

There is a continuum of agents, indexed by $i \in [0, 1]$. Agents are randomly paired to play the following one-shot prisoners' dilemma:

		player j	
		C	D
player i	C	(g, g)	$(-l, g + d)$
	D	$(g + d, -l)$	$(0, 0)$

where C is cooperate or exert effort, D is defect or not exert effort, $i, j \in [0, 1]$. g , l , and d are payoffs for players, where g and l represent gain and loss, respectively, of cooperative behavior, while d is for extra gain from defecting.

To make the game interesting to our purpose, we make the following two assumptions.³

$$d < l, \tag{1}$$

$$g + d - l > 0. \tag{2}$$

The first assumption is crucial to generate reciprocal behavior.⁴ Together with the second one, it guarantees that cooperative behavior always increases total outputs.⁵

2.1 Cooperative Tendency

The utility of player i matched with a partner j is

$$u_i(A_i, A_j) = \underbrace{m_i(A_i, A_j)}_{\text{game-specific}} - \underbrace{\alpha_i \chi_D(A_i)}_{\text{player-specific}},$$

where $A_i, A_j \in \{C, D\}$ are the actions of player i and j , $m_i(A_i, A_j)$ is the material payoff for player i as shown in the above prisoner's dilemma; $\alpha_i \in R^+$ is the disutility player i incurs when defecting, and $\chi_D(A_i)$ is an index function such that

$$\chi_D(A_i) = \begin{cases} 1 & \text{if } A_i = D \\ 0 & \text{if } A_i = C. \end{cases}$$

Note that α_i actually measures player i 's taste for cooperation, or *cooperative tendency*. It can be thought as an internal discipline against defecting, enabling its owner to cooperate in situations where cooperation is otherwise impossible.⁶ Players have heterogenous cooperative tendencies such that $\alpha_i \sim F(\cdot)$, where $F(\cdot)$ is a cdf.

³These assumptions are quite standard in the relevant literature (Kreps et al 1982, Rotemberg 1994, Bar-Gill and Fershtman 2000).

⁴Note that d and l represent a player's marginal costs of cooperating when his partner cooperates and defects, respectively. When a player i plays C , his partner j gets g if playing C , $g + d$ if playing D . The difference of the two payoffs, d , is the marginal cost or net loss of j playing C when i also plays C . Similar arguments apply for l .

⁵Both players exerting effort yields a higher payoff than only one doing so, which is true iff $2g > g + d - l \Leftrightarrow d < l$. Furthermore, exerting effort unilaterally is better than no effort at all, which means $g + d - l > 0$.

⁶This formulation of preferences is supported by the experimental finding that warm-glow effects are highly significant in inducing cooperation in public good games (Palfrey and Prisbrey 1997). People are said to be motivated by warm-glow effects if they derive utility from the very act of cooperating, independent of the exact utility their cooperative behavior delivers to others. In this sense α_i measures the warm-glow motivation of player i .

Each player’s payoff from the game is thus composed of two parts: $m_i(A_i, A_j)$ is game-specific and does not vary across players, while $\alpha_i \chi_D(A_i)$ is player-specific and stable across games. Accordingly, the payoffs associated with the same actions in the above prisoner’s dilemma differ across players. To avoid confusion and be consistent with standard usage in the literature, we call a game a prisoner’s dilemma if it is so for players with zero cooperative tendency.

These two components correspond respectively to two different ways of inducing cooperation in a prisoner’s dilemma. The conventional way is embedding the dilemma in a bigger game to change its game-specific payoffs. For example, appropriate rewards and punishment associated with repeated interactions can transform a stand-alone prisoner’s dilemma to a new one where cooperation becomes a Nash equilibrium. The other way, which this paper adopts, assumes non-standard utility functions where people care about things other than material outputs (e.g. Rotemberg 1994, Bar-Gill and Fershtman 2000). As long as these ‘special’ preferences can be observed and/or intentionally developed in reality, this line of research also yields insightful and refutable results.

2.2 Trustworthiness, Trust, Social Trust

Fix a game γ . Players with different cooperative tendencies can be categorized into three possible behavioral types. We call a player the *selfish* type if he always defects, the *selfless* type if he always cooperates, and the *reciprocal* type if he makes in-kind responses to his partner’s action.⁷ The latter two types are also called *cooperative* or *non-selfish*. The proportions of these types, denoted by $\pi_{S\gamma}$ (selfless) and $\pi_{R\gamma}$ (reciprocal), are determined jointly by the distribution of cooperative tendency and specific game features.⁸

The *trustworthiness* of a player in game γ is the probability that he will cooperate in it. Selfish players have zero trustworthiness since they will never cooperate, while the selfless

⁷Many experimental studies have found that between 40 and 66 percent of subjects exhibit reciprocal behaviors, while between 20 and 30 act completely selfish (Fehr and Gächter 2000).

⁸Various types of cooperative players such as tit-for-tat (Kreps et al. 1982), reciprocal (Fehr and Gächter 2000), and selfless (Tirole 1996) ones are assumed exogenously given in the literature. All of them, however, can be derived from our formulation of cooperative tendency.

ones have full trustworthiness. The trustworthiness of a reciprocal player is zero if matched with a selfish partner, one if a cooperative partner.

How much *trust* a player has in his partner is equal to the latter's trustworthiness. So all players have full trust in a selfless partner, but no trust in a selfish one. A cooperative player will trust a reciprocal partner, but a selfish one not.

The *social trust* in a group is the expected amount of trust one puts in a typical group member. It is determined by the distribution of cooperative tendency $F(\cdot)$ and specific game features. In game γ , it can be characterized by the sufficient statistics $(\pi_{R\gamma}, \pi_{S\gamma})$, the proportions of reciprocal and selfless players.

2.3 Social Trust in Various Games

In the following we study the levels of social trust and its effects on outputs in various games. Specifically we will prove the following proposition.

Proposition 1 *Fix the distribution of cooperative tendency in the population. i) Cooperation exists in one-period prisoner's dilemmas if and only if there is social trust among players. The total outputs strictly increase with social trust. ii) Social trust varies across games. It decreases with d or l or both. iii) In finitely repeated games, the existence of reciprocal players is a necessary condition to generate reputation effects.*

2.3.1 One-period Complete Information Game

Suppose cooperative tendencies are observed publicly. The one-period game between player Ann with cooperative tendency α_A and player Mike with α_M is:

		Mike	
		C	D
Ann	C	(g, g)	$(-l, g + d - \alpha_M)$
	D	$(g + d - \alpha_A, -l)$	$(-\alpha_A, -\alpha_M)$

Claim 1 *In the above one-period complete information game, i) players with cooperative tendencies in the ranges $[0, d)$, $[d, l)$, and $(l, +\infty)$ are of selfish, reciprocal, and selfless type, respectively; ii) (C, C) is a Nash equilibrium if and only if both players are non-selfish.*

Proof. *i)* In this game when Mike plays C , Ann will play C if $g \geq g + d - \alpha_A \Leftrightarrow \alpha_A \geq d$ holds, and she will play D otherwise. When Mike plays D , Ann will play C if $\alpha_A \geq l$ and play D otherwise. In summary, Ann's best response is: always defect when $\alpha_A < d$, always cooperate when $\alpha_A \geq l$, behave reciprocally otherwise. Since the game is symmetric, Mike has the same best response function.

ii) Now we prove that (C, C) is a Nash equilibrium if both players are non-selfish. When two selfless players meet, they always cooperate so that the unique Nash equilibrium is (C, C) . When both players are reciprocal, (C, C) is again a Nash equilibrium since each will play C if the other one plays C .⁹ When a reciprocal player meets with a selfless one, the only Nash equilibrium is also (C, C) since the selfless player always cooperates.

Finally we prove that if either of the two players is selfish, (C, C) is not a Nash equilibrium. This is trivial since a selfish player never plays C . ■

Let π_{RC} denote the proportion of the reciprocal type under complete information, and π_{SC} the selfless type. By definition $\pi_{RC} = \Pr(d \leq \alpha_i < l) = F(l) - F(d)$, $\pi_{SC} = \Pr(\alpha_i \geq l) = 1 - F(l)$. When players are randomly matched with each other, the social trust is π_{SC} from selfish players' perspective, $(\pi_{RC} + \pi_{SC})$ for non-selfish players. It is obvious to see that π_{SC} decreases with l , and $(\pi_{RC} + \pi_{SC})$ decreases with d .

The expected material outputs for each type of players, $\pi_{SC}(g + d)$ for selfish players, $(\pi_{RC} + \pi_{SC})g$ for reciprocal players, and $(\pi_{RC} + \pi_{SC})g - (1 - \pi_{RC} - \pi_{SC})l$ for selfless players, all strictly increase with social trust π_{SC} or $(\pi_{RC} + \pi_{SC})$. So does the total output.

An alternative matching system is assortative matching where selfish players match with each other and cooperative ones match among themselves. In this case, the social trust is zero for selfish players, one for cooperative players, and the average social trust in the group is $(\pi_R + \pi_S)$. All non-selfish players produce output g , while all selfish ones zero. The total output in this case, $Q_1^a \equiv g(\pi_R + \pi_S)$, strictly increases with the average social trust.

⁹Note that (D, D) is another Nash equilibrium when both players are reciprocal. However, each player can unilaterally avoid (D, D) by always playing C since the partner is known to be reciprocal. So individual utility maximization will essentially eliminate (D, D) and leaves the Pareto dominant (C, C) as the only NE ever played between two reciprocal players.

2.3.2 One-period Incomplete Information Game

Under incomplete information players' cooperative tendencies are private information. Let π_{RI} and π_{SI} respectively denote the proportion of reciprocal and selfless type under incomplete information. Let π denote the proportion of cooperative players in equilibrium, i.e. $\pi \equiv \pi_{RI} + \pi_{SI}$.

Claim 2 *In the one-period game under incomplete information, the Bayesian Nash equilibrium is "all players with $\alpha_i \geq \pi d + (1 - \pi)l$ play C, others play D," where π is uniquely determined by the equation $\pi + F(\pi d + (1 - \pi)l) = 1$. Furthermore, $\frac{\partial \pi}{\partial d} < 0$, $\frac{\partial \pi}{\partial l} < 0$.*

Proof. In this game a player's probability of matching with a cooperative partner is believed to be π . By playing C, player i gets g if her partner is cooperative, $-l$ if her partner is defecting. So her expected payoff from playing C is $\pi g - (1 - \pi)l$. By playing D her expected utility is $\pi(g + d - \alpha_i) - (1 - \pi)\alpha_i$. She will play C iff $\pi g - (1 - \pi)l \geq \pi(g + d - \alpha_i) - (1 - \pi)\alpha_i$, which is simplified to $\alpha_i \geq \pi d + (1 - \pi)l$.

To guarantee that the belief π is consistent with players' strategies, it must be true that $\pi = \Pr(\alpha_i \geq \pi d + (1 - \pi)l) \equiv 1 - F(\pi d + (1 - \pi)l)$. The *RHS* is continuous in π on the closed interval $[0, 1]$. It increases with π since $\frac{\partial RHS}{\partial \pi} = (l - d)DF \geq 0$. Furthermore $RHS(\pi = 0) = 1 - F(l) \geq 0$ and $RHS(\pi = 1) = 1 - F(d) \leq 1$, implying that $(l - d)DF < 1$. So π is uniquely determined in the interval $[1 - F(l), 1 - F(d)] \subseteq [0, 1]$.

By the Implicit Function Theorem, $\frac{\partial \pi}{\partial d} = -\frac{\pi d F}{1 - (l - d) d F} < 0$, $\frac{\partial \pi}{\partial l} = -\frac{(1 - \pi) d F}{1 - (l - d) d F} < 0$. ■

Let $\underline{\alpha}$ denote the minimum cooperative tendency to become cooperative, then $\underline{\alpha} \equiv \pi d + (1 - \pi)l$. Claim 2 implies that under incomplete information, players with $\alpha_i < \underline{\alpha}$ are of selfish type, $\alpha_i \in [\underline{\alpha}, l)$ reciprocal, while those with $\alpha_i \geq l$ are again selfless. So we have $\pi_{SI} = 1 - F(l)$, $\pi_{RI} \equiv F(l) - F(\underline{\alpha})$.

Social trust in this game is characterized by $\pi = \pi_{SI} + \pi_{RI} = 1 - F(\underline{\alpha})$, the proportion of cooperative players. Since $l > \underline{\alpha} > d$ for all $\pi \in [0, 1]$, we know $1 - F(l) < 1 - F(\underline{\alpha}) < 1 - F(d)$, which implies that $\pi_{SC} < \pi < \pi_{SC} + \pi_{RC}$. So compared with complete information, social trust under incomplete information is lower for cooperative players, but higher for

selfish players.

The expected output for a selfish player, $G_1^M \equiv \pi(g + d)$, is higher than that of a cooperative player, $G_1^A \equiv \pi g - (1 - \pi)l$. Both, however, strictly increase with social trust π , so does the total output. Note that cooperative players get higher marginal benefit from social trust than that of selfish ones.

2.3.3 T-period Incomplete Information Game

The above analysis has shown that in one-shot games, social trust improves outputs by enabling non-selfish players to cooperate in an otherwise prisoner's dilemma. Now we show that in repeated games social trust can elicit cooperative behavior even from selfish players through reputation effects.

To illustrate the important interaction between social trust and repeated games, we characterize a sequential equilibrium in a finite T-period game under incomplete information. Suppose players are randomly paired to play the above stage game for finite $T \geq 2$ periods. Each pair lasts T periods after they are matched. Their actions are observed at the end of each period. Let $\beta \in [0, 1]$ denote the time discount factor for all players. π and π_{RI} are defined the same as above.

Claim 3 *In this T-period game, the following strategy profile and belief system is a sequential equilibrium if $\beta \geq \frac{d}{(g+d)\pi_{RI}}$ and $\pi_{RI} \geq \frac{d}{g+d}$. The strategy profile is: (1) Selfless players always play C; (2) Reciprocal players play C first; play C if (C, C) is played in the previous period, play D otherwise. (3) Selfish players mimic reciprocal players until period T; play D at period T. The belief system is: (1) In the first period and every period following the history in which only (C, C) has been played, every player assigns probability π to his partner non-selfish. (2) In all the following periods after the first time (C, D) is observed, the player who has played D is believed to be selfish, the player who has played C is still believed to be non-selfish with probability π .*

Proof. In the appendix. ■

In this equilibrium, all players are cooperating except in the last period where players' behaviors are the same as in Claim 2. When we look at each period *isolated*, it seems that repeated interactions *promote* 'trust' among players. But this cannot be reconciled with players' belief that the expected trustworthiness of their partners (and thus the social trust) is always π on the equilibrium path.

Actually when there are no reciprocal players, the reputation effect vanishes and selfish players will not cooperate anymore. So the true motivation for selfish players to cooperate is the reputation effects, the scheme of rewards and punishments made available by repeated interactions and social trust.¹⁰ In comparison, non-selfish types cooperate solely out of their innate discipline, i.e. their cooperative tendencies.

Since many institutions, such as social networks and norms, involve repeated interactions, this example illustrates the general relationship between them and social trust in promoting cooperation.

2.4 Empirical Measures of Social Trust

There are survey-based and experiment-based measures of social trust. And the discrepancies among them are quite common (Glaeser et al. 2000b, Burlando and Hey 1997, Weimann 1994). Now we use the concepts and results developed above to investigate these measures.

The trust question used in the World Values Survey is "Generally speaking, would you say that most people can be trusted, or that you can't be too careful in dealing with people?" The percentage of respondents in each nation replying "most people can be trusted" is the widely used trust indicator, *TRUST*. Under certain conditions, this indicator exactly measures the social trust in a country.

In daily life we often randomly meet each other in some one-period prisoner's dilemma without knowing our partners' individual cooperative tendencies. Suppose in a country C

¹⁰The reciprocal players here act in a similar way as tit-for-tat players in Kreps et al. (1982). However, they are always rational utility-maximizers, while tit-for-tat players rigidly follow a certain behavior rule regardless of its associated utility. So the formulation of cooperative tendency is more general than tit-for-tat preference.

the representative dilemma is γ_C , and the proportion of cooperative players is π_C . When these people are asked the same trust question as in the World Values Survey, what is the percentage replying “most people can be trusted”? It is π_C , since exactly π_C proportion of players meet a partner that can be trusted. So we get $TRUST_C = \pi_C$.

In a public goods experiment γ_P , suppose the proportion of cooperative players (subjects who make positive contributions) is π_P . If the distribution of cooperative tendency among the subjects participating γ_P is a random sample drawn from the whole population, and γ_P is the same as γ_C , then π_P is an unbiased estimate of $TRUST$. Otherwise, discrepancies among different measures of social trust arise.

3 Social Trust Formation

We have shown that social trust is crucial to induce cooperation in a society. This begs the following important question: how social trust is formed? Though little about the social psychological process of cooperative tendency development is known, social learning model is suggested to be an appropriate proxy.¹¹ For example, parents and teachers, acting as role models and choosing appropriate home and school inputs, may deliberately teach children to be more cooperative.¹²

Suppose parents are able to invest in their children positive cooperative tendency with some costs. The question is, would they choose to do so in equilibrium? If so, what is the equilibrium social trust level, is it optimal? How can we improve social trust? These issues are addressed in this section.

3.1 The Basic Model

Each player lives two periods. The first period is the investment stage where each player’s cooperative tendency is chosen (by their parents) to maximize his life-time utility, taking as

¹¹See Cook (2001) and Cappelli (1995) for some evidence.

¹²Indeed, children’s cognitive and social development are affected by home inputs (Huang 2002). And parents do choose certain desirable traits to invest in children. For example, 77.2% of parents think that “help others when they need help” is one of the three most important traits that their children should learn, and 96.8% rank it among the top four (General Social Survey from 1986 to 1998).

given the expected proportion of cooperative players $\Pi \in [0, 1]$ in the population. Investing in cooperative tendency incurs positive cost, and the cost is higher for players with higher index. In particular, the cost function is $c(\alpha, i)$, where $\alpha \in R^+$, $i \in [0, 1]$. We assume $c(0, i) = 0$, $c_\alpha > 0$, $c_i > 0$, $c_{\alpha\alpha} \geq 0$, $c_{\alpha i} \geq 0$.

The second period is the production stage. With probability $1 - p$, players' cooperative tendencies are private information and they randomly match each other to play the one-period prisoner's dilemma characterized by (g, d, l) . With probability p , players' cooperative tendencies are publicly observed and they are free to choose their partners, playing a one-period prisoner's dilemma characterized by (G, D, L) , where $G \geq g$ and $D \geq \Pi d + (1 - \Pi)l$.

In this environment, players either do not invest at all in cooperative tendency, or if they ever invest, their cooperative tendency will be equal to D , just enabling them to cooperate in the complete information game.

Lemma 1 $\alpha_i = D$ iff $\alpha_i > 0$ for any player i .

Proof. Players with $\alpha \geq D$ will cooperate in both games and get utility $pG + (1 - p)[\Pi g - (1 - \Pi)l]$, independent of α . Since investing in α is costly, they will choose the lowest possible level D . Players with $\alpha < \Pi d + (1 - \Pi)l$ will always defect and get $(1 - p)\Pi(g + d) - p\alpha$. Their payoff is maximized when $\alpha = 0$. Any cooperative tendency in the middle range will make players worse off than otherwise. The reason is that these α will enable players to cooperate in game (g, d, l) but not in (G, D, L) , getting payoff $(1 - p)[\Pi g - (1 - \Pi)l] - p\alpha$ that is always lower than those of the other two groups. ■

Without much loss of generality, we assume $D = \Pi d + (1 - \Pi)l$. Let $V(D, i)$ denote the expected life-time utility for player i when he becomes cooperative, and $V(0, i)$ if otherwise.

$$\begin{aligned} V(D, i) &= \beta p G + \beta(1 - p)[\Pi g - (1 - \Pi)l] - c(D, i), \\ V(0, i) &= \beta(1 - p)\Pi(g + d). \end{aligned}$$

Let $V_d(i, \Pi)$ represent the net return of investing in cooperative tendency v.s. remaining

selfish. Players will choose to invest if and only if $V_d \geq 0$. By definition,

$$\begin{aligned} V_d(i, \Pi) &\equiv V(D, i) - V(0, i) \\ &= \beta p G - \beta(1-p)[\Pi d + (1-\Pi)l] - c(D, i). \end{aligned}$$

The following lemma says $V_d(i, \Pi)$ increases with the expected social trust Π and decreases with a player's index.

Lemma 2 $\frac{\partial V_d(i, \Pi)}{\partial i} < 0$, $\frac{\partial V_d(i, \Pi)}{\partial \Pi} > 0$.

Proof. $\frac{\partial V_d(i, \Pi)}{\partial \Pi} = \beta(1-p)(l-d) + c_D(l-d) = \beta(1-p+c_D)(l-d) > 0$.

$\frac{\partial V_d(i, \Pi)}{\partial i} = -c_i < 0$. ■

The intuition is quite clear. A marginal increase in Π not only improves the chance of meeting a cooperative player, but also reduces the cost of investing in appropriate cooperative tendency. Since cooperative players benefit more from both channels, the return of being cooperative strictly increases with Π . $V_d(i, \Pi)$ decreases with players' index because the investing cost increases with index.

3.2 Positive Externalities

Suppose the social planner's objective function is to maximize the sum of all players' lifetime utility:

$$\max_{\pi} V(\pi) = \int_{i=0}^{i^S} V(D, i) di + \int_{i=i^S}^1 V(0, i) di,$$

where π is the proportion of cooperative players, and i^S is the highest index among them.

Then $\pi = \Pr(i \leq i^S) = i^S$, since i is uniformly distributed on the interval $[0, 1]$.¹³

Proposition 2 *The social returns of investment in cooperative tendency are strictly larger than individual returns.*

Proof. Take the first derivative of social welfare $V(\pi)$ with respect to π , we get

$$\frac{dV}{d\pi} = \underbrace{\int_{i=0}^{i^S} \frac{\partial V(D, i)}{\partial \pi} di + \int_{i=i^S}^1 \frac{\partial V(0, i)}{\partial \pi} di}_{\text{externality on others due to } \pi \text{ increase}} + \underbrace{[V(D, i^S) - V(0, i^S)]}_{\text{individual return for player } i^S}.$$

¹³A general distribution function also works.

The social return is composed of two parts: the individual return for player i^S who becomes non-selfish, and the externality on all other players due to the marginal increase of π . The externality is positive because all players benefit from social trust increase: $\frac{\partial V(D,i)}{\partial \pi} = \beta(1-p)(g+l) + c_D(l-d) > 0$, $\frac{\partial V(0,i)}{\partial \pi} = \beta(1-p)(g+d) > 0$. So the social return for any player being non-selfish is always strictly bigger than his individual return. ■

This proposition implies that the equilibrium social trust level is always strictly lower than the socially optimal level, except when there is already full trust in equilibrium. Equivalently, the investment in cooperative tendency is generally not efficient.¹⁴

3.3 The Equilibrium

Now we study the existence and properties of Nash Equilibrium (*NE* thereafter) at the investment stage. Note that every *NE* can be characterized by a pair $(\Pi = e, \pi = e)$, where π is the actual proportion of non-selfish players, and $e \in [0, 1]$.¹⁵ And the ‘no social trust’ equilibrium $(\Pi = 0, \pi = 0)$ always exists regardless of the underlying fundamentals.¹⁶

We partition the parameter space into four cases and characterize the corresponding equilibria. We further consider these *NEs* as steady states in a dynamic process to check whether they are stable to small perturbations of Π .¹⁷

3.3.1 The Benchmark Case

In this case the net returns of investing in cooperative tendency are quite similar across players, not too high or too low. Specifically, there exist $\Pi_0, \Pi_1 \in (0, 1)$ such that player 0 is indifferent to investing or not when $\Pi = \Pi_0$, and player 1 indifferent when $\Pi = \Pi_1$. Thus

¹⁴Though it is very probable that everybody becoming cooperative is the social optimal solution, this may not always be the case. For example, if the investing costs for some players are extremely high, say higher than the positive externalities received by all other players, then it is better that they remain selfish.

¹⁵Given all other players’ strategies (summarized by Π), player i invests in α if and only if $V_d(i, \Pi) \geq 0$. No player wants to deviate from this choice when the expected social trust is exactly realized, i.e. when $\pi = \Pi$.

¹⁶The proof is trivial since there is no gain from being the only cooperative player.

¹⁷A dynamic process is like this. There are countable infinite generations of players denoted by $1, 2, \dots, N, \dots$. Each generation is identical. The expected social trust in every following generation is equal to the realized social trust in its immediately previous generation. That is, $\Pi_{N+1} = \pi_N, \forall N = 1, 2, \dots$. The initial one, $\Pi_{N=1}$, is assumed exogenously given.

the following two conditions

$$V_d(0, \Pi_0) = 0, \quad (3)$$

$$V_d(1, \Pi_1) = 0, \quad (4)$$

characterize the benchmark case.

Lemma 3 $\Pi_0 < \Pi_1$.

Proof. By Lemma 2 $\partial V_d(i, \Pi)/\partial i < 0$. So $V_d(1, \Pi_0) < V_d(0, \Pi_0) = 0$. We also have $V_d(1, \Pi_1) = 0$ by condition (4). Combining them we get $V_d(1, \Pi_0) < V_d(1, \Pi_1)$. But we know $\partial V_d(i, \Pi)/\partial \Pi > 0$ by Lemma 2, so $\Pi_0 < \Pi_1$. ■

Lemma 4 *i) The best response function is*

$$B(\Pi) \equiv \begin{cases} 0 & \text{if } \Pi \in [0, \Pi_0] \\ i^*(\Pi) & \text{if } \Pi \in [\Pi_0, \Pi_1] \\ 1 & \text{if } \Pi \in [\Pi_1, 1] \end{cases},$$

where $i^*(\Pi)$ is the unique solution to $V_d(i(\Pi), \Pi) = 0$ for any $\Pi \in [\Pi_0, \Pi_1]$.

ii) $B(\Pi)$ is continuous, strictly increasing in Π on $[\Pi_0, \Pi_1]$, and weakly increasing on $[0, 1]$.

Proof. *i)* By Lemma 2, $V_d(i, \Pi)$ is continuous and strictly decreasing in $i \in [0, 1]$ for any Π . From Lemma 3 we know $V_d(0, \Pi) > 0$ and $V_d(1, \Pi) < 0$, for any $\Pi \in [\Pi_0, \Pi_1]$. These two conditions guarantee that for each $\Pi \in [\Pi_0, \Pi_1]$, there exists a unique $i^* \equiv i^*(\Pi) \in [0, 1]$ such that

$$V_d(i^*(\Pi), \Pi) = 0.$$

This means that all players with lower index than $i^*(\Pi)$ will choose to become cooperative and others not, so that $B(\Pi) = \Pr(i \leq i^*(\Pi))$. Since i is uniformly distributed on $[0, 1]$, we have $B(\Pi) = i^*(\Pi)$.

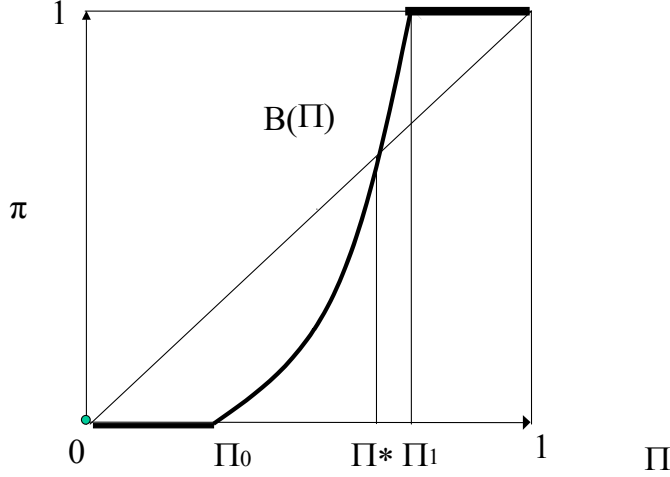
Using similar arguments, it is trivial to show that $B(\Pi) = 0$ for any $\Pi \in [0, \Pi_0]$, and $B(\Pi) = 1$ for any $\Pi \in [\Pi_1, 1]$.

ii) $B(\Pi)$ strictly increases in Π on $[\Pi_0, \Pi_1]$ since

$$\frac{\partial i^*(\Pi)}{\partial \Pi} = -\frac{\partial V_d(i^*, \Pi)/\partial \Pi}{\partial V_d(i^*, \Pi)/\partial i^*} > 0.$$

It is then trivial to verify that $B(\Pi)$ is continuous and weakly increasing in Π on $[0, 1]$. ■

Since $B(\Pi)$ is continuous and strictly increasing in Π on $[\Pi_0, \Pi_1]$, and $B(\Pi_0) = 0$, $B(\Pi_1) = 1$, there must exist at least one fixed point $\Pi^* \in [\Pi_0, \Pi_1]$ such that $i^*(\Pi^*) = \Pi^*$.¹⁸ When $B(\Pi)$ has monotone slopes on the interval $[\Pi_0, \Pi_1]$, an assumption we will maintain in this section, the *NE* $(\Pi = \Pi^*, \pi = \Pi^*)$ is unique.¹⁹ It is easy to check that $(\Pi = 0, \pi = 0)$ and $(\Pi = 1, \pi = 1)$ are the other two *NEs*. See the following figure for illustration.



Case I

If the initial belief is $\frac{\varepsilon}{2}$ lower than Π^* , ultimately this economy will fall into the ‘no-trust’ trap $(\Pi = 0, \pi = 0)$. On the contrary, if the initial belief is $\frac{\varepsilon}{2}$ higher than Π^* , the economy will gradually reach ‘full trust’ state $(\Pi = 1, \pi = 1)$. These two corner *NEs* are stable with respect to small perturbations. The interior *NE* (Π^*, Π^*) is unstable, happening only when

¹⁸The exact number of fixed points on $[\Pi_0, \Pi_1]$ depends on the curvature of $i^*(\Pi)$, which is difficult to pin down in general. Note that $\frac{\partial^2 i^*(\Pi)}{\partial \Pi^2} = \frac{\beta(l-d)^2[(1-p+c_D)c_{iD}-c_{DD}c_i]}{c_i^2}$ is positive if $c_{DD} = 0$ or $c(D, i) = D^2 + Di$.

¹⁹For example, when the best response function is linear, it must be $i^*(\Pi) = a\Pi - b$, where $a = \frac{1}{\Pi_1 - \Pi_0}$ and $b = \frac{\Pi_0}{\Pi_1 - \Pi_0}$ are determined by conditions (3) and (4). Note that the slope a is bigger than 1. Thus we get $i^*(\Pi) = \frac{\Pi - \Pi_0}{\Pi_1 - \Pi_0}$. Let Π_l^* be the solution to $i^*(\Pi_l^*) = \Pi_l^*$, then $\Pi_l^* = \frac{\Pi_0}{1 + \Pi_0 - \Pi_1}$. It is trivial to check that $\Pi_l^* \in [\Pi_0, \Pi_1]$.

the initial belief is exactly Π^* . Thus we have proved the following proposition.

Proposition 3 *Under conditions (3) and (4), there are three NEs: $(0, 0)$, (Π^*, Π^*) , and $(1, 1)$, where $\Pi^* \in [\Pi_0, \Pi_1] \subset (0, 1)$. Among them $(0, 0)$ and $(1, 1)$ are stable.*

This case shows that a negligible ε difference in initial beliefs can lead to two polar stable equilibria. It may account for the existence of dramatically different performances between two otherwise identical communities or organizations (Putnam 1993). The intuition is as follows. Players' investment costs in this case are quite tightly distributed in the middle range, so that other people's choices are relatively more important in affecting players' investment decision than their idiosyncratic differences in cost. If they believe that enough people (over the threshold Π^*) plan to invest in cooperative tendency, then the expected return is positive to everybody so that all will jump into the bandwagon, and vice versa. In other words, nobody is different enough in their investing costs to avoid being swept away by others' choices.

3.3.2 Diverse Cost Case

In contrast to the benchmark case, players here have quite diverse costs. Some players have costs so low that they want to invest in cooperative tendency no matter how few players, as long as more than zero, are expected to do so. On the other hand, there are players whose costs are so high that they choose not to invest even when everybody else is expected to do so. This case is characterized by the following two conditions.

$$\lim_{\Pi \rightarrow 0^+} V_d(0, \Pi) > 0. \quad (5)$$

$$V_d(1, 1) < 0. \quad (6)$$

Condition (5) is equivalent to $\pi_0 > 0$, where π_0 is defined by $\lim_{\Pi \rightarrow 0^+} V_d(i^*(\Pi), 0) = 0$ and $\pi_0 \equiv \lim_{\Pi \rightarrow 0^+} i^*(\Pi)$. Similarly, condition (6) can be written as $\pi_1 < 1$, where π_1 is determined by $V_d(i^*(1), 1) = 0$ and $\pi_1 \equiv i^*(1)$. These two conditions imply that, as long as the expected social trust is positive, regardless of how close it might be to zero, there are

π_0 players choosing to be cooperative; on the other hand, there are at most π_1 cooperative players when the expected social trust is one.

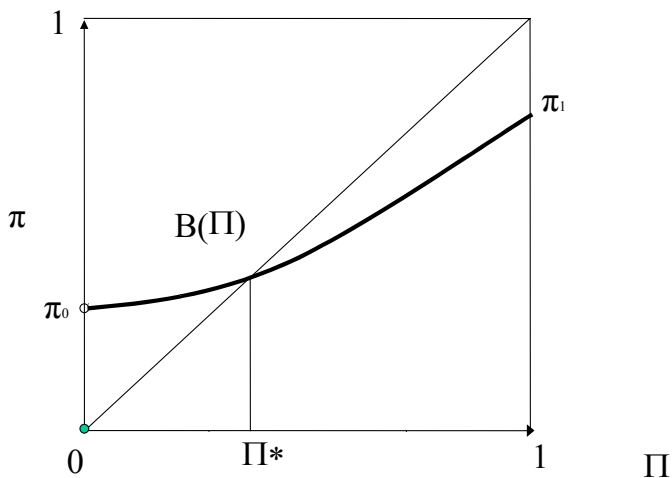
Proposition 4 *Under conditions (5) and (6), there exist two NEs: $(0, 0)$ and (Π^*, Π^*) , where $\Pi^* \in (\pi_0, \pi_1)$. Only (Π^*, Π^*) is stable.*

Proof. Similar arguments as in the benchmark case lead to the best response function²⁰

$$B(\Pi) = \begin{cases} 0 & \text{if } \Pi = 0 \\ i^*(\Pi) & \text{if } \Pi \in (0, 1] \end{cases}$$

Let $\Pi^* \in (0, 1)$ denote the solution to the equation $i^*(\Pi^*) = \Pi^*$, then $\Pi^* \in (\pi_0, \pi_1)$ because $B(\Pi \rightarrow 0) = \pi_0$, $B(\Pi = 1) = \pi_1$, and $B(\Pi)$ strictly increases in $\Pi \in (0, 1]$. ■

Here the interior NE (Π^*, Π^*) is the only focal point of the history, immune to random events. The reason is contrary to the benchmark case. Here players are sufficiently different in their investment costs so that the investment externality among players is less important in affecting their choices. Since some players will never invest in cooperative tendency, social optimum is not achievable in equilibrium. See the following figure for illustration.



Case II

²⁰Suppose the linear best response function is $i^*(\Pi) = c\Pi + d$. Then $d = \pi_0$, $c + d = \pi_1$. Let $i^*(\Pi_l^*) = \Pi_l^*$, we get $\Pi_l^* = \frac{\pi_0}{1 + \pi_0 - \pi_1}$.

The following comparative statics suggest that long-run social trust increases with p , β and G , decreases with d and l . These results are also true in stable equilibrium at interior points in other cases.

Proposition 5 $\partial\Pi^*/\partial p > 0$, $\partial\Pi^*/\partial\beta > 0$, $\partial\Pi^*/\partial G > 0$; $\partial\Pi^*/\partial d < 0$, $\partial\Pi^*/\partial l < 0$.

Proof. To prove $\partial\Pi^*/\partial p > 0$, we show that the best response function $B(\Pi)$ is shifted up by p for each $\Pi \in (0, 1]$, and π_0 also increases with p . Accordingly, the intersection of $B(\Pi)$ with the 45⁰ line, Π^* , must also increase with p .

$$\frac{\partial B(\Pi)}{\partial p} = \frac{\partial i^*(\Pi)}{\partial p} = -\frac{\partial V_d(i^*, \Pi)/\partial p}{\partial V_d(i^*, \Pi)/\partial i^*} = \frac{\beta[G + \Pi d + (1 - \Pi)l]}{-\partial V_d(i^*, \Pi)/\partial i^*} > 0,$$

By definition of π_0 , we know

$$\lim_{\Pi \rightarrow 0} V_d(i = \pi_0, \Pi) = \beta p G - \beta(1 - p)l - c(l, \pi_0) = 0.$$

By Implicit Function Theorem, we get $\frac{\partial \pi_0}{\partial p} = -\frac{\beta(G+l)}{-c_i} > 0$.

The other four comparative statics are proved in exactly the same way and thus relegated to the appendix. ■

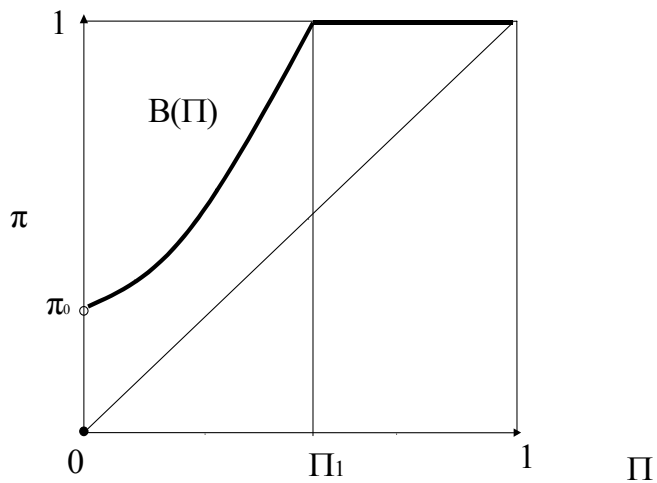
3.3.3 Low Cost and High Cost Cases

Conditions (4) and (5) characterize the low cost case, where $\Pi_1 \in [0, 1]$. Here even the highest cost players invest in cooperative tendency when they believe enough people are doing so. The final case is defined by conditions (3) and (6), where $\Pi_0 \in (0, 1]$. In contrast with the low cost case, here the investing cost is high for all players. See the following figures for illustration.

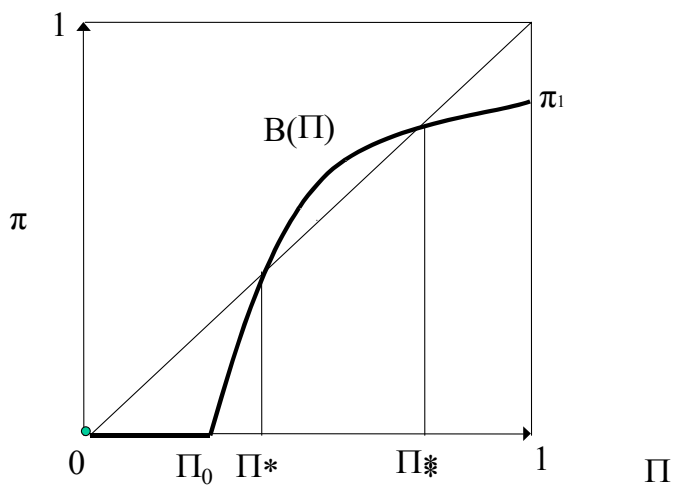
Proposition 6 *i) Under conditions (4) and (5), (1, 1) always exists and is stable. It is either the only equilibrium, or two other NEs exist at interior points where the one with lower social trust is stable.*

ii) Under conditions (3) and (6), (0, 0) always exists and is stable. It is either the only equilibrium, or two NEs exist at interior points where the one with higher social trust is stable.

Proof. *i)* The sufficient condition for unique NE is $\partial^2 B(\Pi)/\partial \Pi^2 \leq 0$, or when $\partial^2 B(\Pi)/\partial \Pi^2 > 0$ and $\Pi_1 \leq \pi_0$ both hold. *ii)* A sufficient condition for unique NE is $\partial^2 B(\Pi)/\partial \Pi^2 \geq 0$. Another one is $\partial^2 B(\Pi)/\partial \Pi^2 < 0$ and $\Pi_0 \leq \pi_1$. The proof is otherwise similar to the other two cases and thus omitted. ■



Case III



Case IV

Proposition 7 *i) Full trust is achievable in stable NE when condition (4) is satisfied but never so when (6) holds. ii) ‘No trust’ equilibrium always exists. It is stable under condition (3). iii) Multiple equilibrium is possible in all cases. However, among all NEs with social trust levels in $(0, 1)$, the stable one is unique.*

The above proposition summarizes some common results of the above four cases. It implies that high-cost players are crucial in achieving full trust, while low-cost players are important in generating positive social trust. There exist multiple equilibria where up to two of them are stable. However, there is only one stable equilibrium with $\pi \in (0, 1)$. Our discussions below thus focus on the unique stable equilibrium with non-extreme social trust levels, where the comparative statics are given in Proposition 5.

3.4 Several Ways to Increase Social Trust

3.4.1 The Information Structure

In reality an individual’s trustworthiness sometimes can be assessed from reading his appearance, attitudes, and spontaneous responses. The accuracy of this encoding process often increases with one’s sophistication, and decreases with the difference between partners in backgrounds.²¹ People’s cooperative tendencies are also revealed by their actions in one-period prisoner’s dilemmas according to our results in the second section. Better information flow through efficient mass communication and dense social networks helps facilitate the revelation of cooperative tendencies, since more information can be accumulated about how to assess people’s trustworthiness in certain circumstances, and about particular individuals’ behaviors in various one-period prisoner’s dilemmas.

In our simple social trust formation model the information structure in a society is represented by p , the probability that individual cooperative tendencies are observed. The model (Proposition 5) suggests that social trust in stable equilibrium increases with p . The intuition is that, as the information flow becomes more efficient people have more chances

²¹Indeed, subjects paired with a partner of a different race or nationality are less cooperative in the public goods experiments (Glaeser et al. 2000).

to form cooperative relationships and get benefits, so that they are more willing to invest in cooperative tendency.

This is consistent with the empirical evidence that social trust is positively correlated with both daily newspaper circulation (0.73) and the number of radios per capita (0.53) across 29 countries (Temple and Johnson 1998). They conclude that “an assessment of mass communications, given the absence of other good measures, is probably the best way of capturing variation in social trust across developing countries.” It is helpful to note that social networks are quite dense in developing countries.

In developed countries the information structure is quite different. Mass communication is quite efficient everywhere due to the advancement in information technology, but the density of social networks is weakened since high mobility rate increases background variation in one’s acquaintances, making it more costly to build up social networks. This latter force may have contributed to the steady decline of trust indicators in US (Putnam 1995).

3.4.2 Extrinsic incentives and intrinsic Discipline

Material payoffs in the prisoner’s dilemmas, such as d and l , represent *extrinsic incentives* to defect. They are determined by institutions including the legal system, incentive schemes in organizations, social networks, and social norms. The more effective these institutions in punishing defecting behaviors, the lower d and l .

In contrast, cooperative tendency is an *intrinsic discipline* against defecting. The process of developing cooperative tendency in children is primarily conducted at home and in schools. The lower the investment cost, the higher people’s cooperative tendencies.²²

Both extrinsic incentives and intrinsic discipline can affect people’s behaviors. To achieve cooperation we can either improve the efficiency of institutions, or reducing the investing cost of cooperative tendency, or both. How to allocate resources between them depends on their relative costs. When cooperative tendency is endogenous, however, their relationship

²²For example, children of poor family backgrounds who attended early intervention programs like Head-start are more likely than others to adopt pro-social behaviors (Heckman 1999).

is more complex than simple substitution.

Proposition 8 *Effective disciplinary institutions increase social trust in the context of these institutions, but reduce the average cooperative tendency.*

Proof. Effective disciplinary institutions reduce d and l , the extrinsic incentives to defect. By Proposition 5 we know $\frac{\partial \Pi^*}{\partial d} < 0$, $\frac{\partial \Pi^*}{\partial l} < 0$. That is, social trust in stable equilibrium is higher when d and l are lower. The average cooperative tendency $\underline{\alpha}$, however, decreases with d and l because

$$\begin{aligned}\frac{\partial \underline{\alpha}}{\partial d} &= \Pi^* + (d - l) \frac{\partial \Pi^*}{\partial d} > 0, \\ \frac{\partial \underline{\alpha}}{\partial l} &= (1 - \Pi^*) + (d - l) \frac{\partial \Pi^*}{\partial l} > 0.\end{aligned}$$

Note that the average cooperative tendency is equal to $\underline{\alpha} = \Pi^*d + (1 - \Pi^*)l$. ■

Here the outside discipline crowds out innate discipline in that effective disciplinary institutions reduce defecting benefits and thus the required cooperative tendencies to achieve cooperation.²³ On the other hand, more people can afford the investment costs of lower cooperative tendencies so that social trust is higher. In contrast, when disciplinary institutions are less efficient, people have to invest in higher cooperative tendencies to achieve cooperation. As a result fewer people are cooperative, but their average cooperative tendency is higher.

An interesting implication is that, survey-based social trust measure such as *TRUST* is higher in a country with effective disciplinary institutions, but experiment-based social trust measure may be lower if the game has quite high defecting benefits. This may account for the contradictory social trust ranking across countries using different measures. For example, *TRUST* in UK (44.4) is much higher than that of Italy (26.3), and US (45.4) much higher than Germany (29.8) (Knack and Keefer 1997). However, UK subjects “free-rode to a much greater extent” than Italians in a public goods experiment (Burlando and Hey 1997), and US subjects free-rode more than German ones (Weimann 1994).

²³See Bar-Gill and Fershtman (2000) for a similar situation.

3.4.3 Other Elements

Many elements increase the benefits of investing in cooperative tendency, thus also improves social trust. Two immediate examples are higher time preference β and longer living periods to collect the returns, implying that more patient, younger people are more willing to invest in cooperative tendency.

In the production stage, as the tenure of the established cooperative matches goes up, cooperative players' gain from complete information games becomes bigger. This has the same positive effect on social trust as a larger G in our model. Thus social trust is higher in a society where members of communities and organizations have longer tenures together. For example, when the divorce rate is lower and families are more stable, the average turnover rate within a firm is lower, people are encouraged to become home-owners rather than tenants, the returns to being cooperative are higher. Shorter tenures in families, communities, and firms may also contribute to the decline of social trust in the US.

3.5 An Extension with Human Capital

Cooperative tendency, as a trait invested in a person that yields returns to him/her in the future, is essentially a component of human capital.²⁴ Cognitive ability, the conventional component of human capital denoted by h , directly enters a specific production function. In contrast, cooperative tendency enables people to cooperate with each other, thus increases producing opportunities for h . These two components together determine a person's overall productivity when other things are given.²⁵ At investment stage, their relationship is the same as that between child cognitive and social development.

Now we extend the basic social trust formation model to include cognitive ability. Since

²⁴In the same spirit, some personal characteristics such as working attitude, self-discipline, motivation, and time preference are treated as components of human capital in Becker (1996), Bowles and Gintis (1998), Heckman (2000), and Bowles et al. (2001).

²⁵At the aggregate level, the distribution of cooperative tendency in the population forms the cooperative infrastructure of a society, which determines social trust. Similarly the distribution of cognitive ability in the society is its intellectual infrastructure, its importance in economic growth has been shown by Lucas (1988) among others.

almost all previous results carry over, we only discuss the differences and new findings. The proof for the human capital version of Lemma 2 is in the appendix.

3.5.1 Human Capital Version of the Stage Game

Now we let the material payoffs of player i depend on his human capital (h^i, α^i) . The game γ_h , the human capital version of the prisoner's dilemma between players i and j , is

		Player j	
		C	D
Player i	C	$g(h^i), g(h^j)$	$-l(h^i), g(h^j) + d(h^j) - \alpha^j$
	D	$g(h^i) + d(h^i) - \alpha^i, -l(h^j)$	$-\alpha^i, -\alpha^j$

The production functions $g(\cdot), d(\cdot), l(\cdot)$ are increasing and concave in h , where $d(h) < l(h)$ and $g(h) + d(h) - l(h) > 0$ for all h , corresponding to assumptions (1) and (2). When both players defect they produce the default amount which is again normalized to zero. This implies that all outputs in this game are at relative level, indicating the importance of cooperation in improving productivity.

Under complete information, player i is of selfish type iff $\alpha^i < d(h^i)$, selfless iff $\alpha^i \geq l(h^i)$, reciprocal iff $\alpha^i \in [d(h^i), l(h^i)]$. Under incomplete information, player i cooperates iff $\alpha_i \geq \underline{\alpha}(h^i, \Pi)$, where $\underline{\alpha}(h^i, \Pi) \equiv \Pi d(h^i) + (1 - \Pi)l(h^i)$. Here again Π denotes the expected proportion of cooperative players in the population. These two results are direct extensions of Claim 1 and 2, respectively. A new fact is that cooperative players now have different cooperative tendencies that increase with their cognitive ability h , i.e. $\partial \underline{\alpha}(h^i, \Pi) / \partial h^i \geq 0$. The rationale is that players with higher h are usually faced with greater temptation to defect.

3.5.2 Human Capital Investment Model

This human capital investment model has the same timing and information structure as in the basic model, except that now players have to choose (h, α) together. The cost function is $c(h, \alpha, i)$, where $h, \alpha \in \mathbb{R}^+$, $i \in [0, 1]$, $c(0, 0, i) = 0$, $c_h > 0$, $c_\alpha > 0$, $c_i > 0$, $c_{hh} \geq 0$, $c_{\alpha\alpha} \geq 0$, $c_{\alpha i} > 0$.

Let $V_A^i(h)$ denote the expected life-time utility for player i when he becomes cooperative, and $V_M^i(h)$ if otherwise. We have

$$\begin{aligned} V_A^i(h) &= \beta p G(h) + \beta(1-p)[\Pi g(h) - (1-\Pi)l(h)] - c(h, \underline{\alpha}(h, \Pi), i), \\ V_M^i(h) &= \beta(1-p)\Pi(g(h) + d(h)) - c(h, 0, i). \end{aligned}$$

Let $h_A^i \equiv h_A(\Pi, p, \beta, i, k)$ and $h_M^i \equiv h_M(\Pi, p, \beta, i, k)$ denote the solutions that maximize $V_A^i(h)$ and $V_M^i(h)$ respectively, where k represents all other parameters. Their existence and relevant comparative statics are summarized in the following proposition.

Proposition 9 h_A^i and h_M^i exist and are unique. h_M^i increases in Π and β , but decreases in p . h_A^i increases with p and β . It increases with Π if $c_{h\alpha}(h, \alpha(h, \Pi), i) \geq 0$ and $d'(h) \leq l'(h)$.

Proof. See the Appendix. ■

This proposition suggests that a player will choose different levels of cognitive ability h_A^i and h_M^i corresponding to his choices of cooperative tendency. A cooperative player's cognitive ability h_A^i increases as the probability of complete information p goes up, while the opposite is true for a selfish player. Cognitive ability of both types increases with time discount factor β and expected social trust Π .

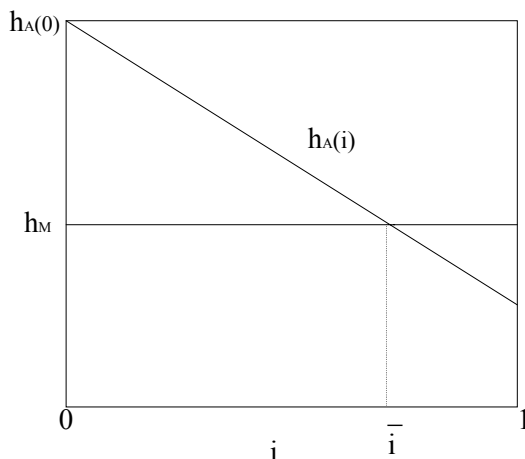
To study the effects of investing in cooperative tendency on cognitive ability among otherwise homogenous players, the marginal cost of investing in h is assumed to be the same for all players, i.e. $c_{hi}(h, \alpha, i) = 0$. Meanwhile recall that $c_{\alpha i}(h, \alpha(h), i) > 0$.

Proposition 10 *i) For any given Π , h_A^i strictly decreases with i , while $h_M^i = h_M$ across all players. ii) There exists a unique $\bar{i}(\Pi, p) \in [0, 1]$ such that $h_A^i \geq h_M$ for all $i \leq \bar{i}$, while $h_A^i < h_M$ for all $i > \bar{i}$. $\bar{i}(\Pi, p)$ increases with p and Π .*

Proof. See the appendix. ■

Note that the optimal level of cognitive ability is affected by investing costs of both components of human capital. Players will choose the same h_M if not investing in cooperative tendency, since their marginal costs are the same. If players invest in cooperative tendency,

however, those with lower investment costs ($i \leq \bar{i}$) may choose higher cognitive ability than h_M , and the opposite is true for high cost players $i > \bar{i}$. Therefore cognitive ability and cooperative tendency complement each other for low cost players, but are substitutes for high cost players. As p or Π goes up, more people will enjoy complementariness between these two components. See the following figure for illustration.



Relation Between h_A^i and h_M

4 Conclusions

Social trust is an important social phenomenon. It is also one of the most important forms of social capital that attracts lot of academic and popular attention. The recent empirical literature has shown that it facilitates economic performance at various levels. The formal analysis of social trust, however, has been lagging behind primarily due to the lack of operational concepts.

This paper formalizes the concept of trust, trustworthiness, and social trust, using a single analytical element ‘cooperative tendency’. In a given game, people with different levels of cooperative tendency have three types of behaviors: selfish, reciprocal, and selfless. Across games, the same person may behave differently, while those with higher cooperative tendencies are more likely to cooperate. These behavioral predictions are consistent with the findings of many public goods experiments.

The amount of social trust is determined by the distribution of cooperative tendency among players and the specific game features. It thus varies across games and players, which explains why there are discrepancies among survey-based trust indicator and social trust measures in various public goods experiments.

The paper shows that cooperation in one-period prisoner's dilemma exists if and only if there is social trust among players. Higher cooperation level in finitely repeated games is caused not by newly generated trust, but by reputation effects. And reputation effects vanish when there is no social trust. The total outputs strictly increase with social trusts, which accounts for the empirical findings that social trust improves economic performance in organizations and at the overall level.

Since appropriate cooperative tendencies enable players to cooperate and produce more outputs, players may want to invest in it to maximize their expected life-time earnings. This investment, however, is in general not efficient because the social returns are always strictly larger than individual returns. Thus social trust in equilibrium is usually lower than the optimal level. Multiple equilibrium always exists, while the stable equilibrium with non-extreme social trust is unique. The equilibrium social trust can be improved by increasing the efficiency of information flowing, disciplinary institutions, and the process of investing in cooperative tendency in a society.

The cooperative tendency is in essence a component of human capital, distinct from cognitive ability. Both are costly to invest and can generate returns in the future. For example, parents often have to optimally balance a child's social and cognitive development. Under certain conditions, however, these two components of human capital actually complement each other. So investing in cooperative tendency creates a new margin to boost cognitive ability development.

The social trust formation model clarifies the mechanism of various forces in improving social trust. In particular, it shows that a society's institutions, acting as outside disciplines, increase equilibrium social trust level, but may crowd out individuals' internal discipline. In contrast, reducing the cost of investing in cooperative tendency at home and schools

increases both social trust and the average cooperative tendency in the society.

In economics literature, much attention has been focused on outside disciplinary institutions, while the importance of using inside discipline to improve social trust is not well appreciated. This bias is also prominent in many real life situations. For example, the current policies regarding education and job training “...focus on cognitive skills ... to the exclusion of social skills, self-discipline and a variety of non-cognitive skills that are known to determine success in life” (Heckman 2000). Not coincidentally, firms often report shortage in appropriate working habits and attitudes, suggesting a severe under-investment (Cappelli 1995).

There are many possible reasons suggested by the paper to account for these phenomena. First, individuals lack appropriate incentives to develop cooperative tendencies due to strong positive externalities (Proposition 2). Second, the investment cost is high. The process of nurturing cooperative tendency starts from early childhood and continues almost to or beyond adulthood, which involves a very long time. And the specific technology is not well understood and mastered yet. Third, people may not fully recognize the importance of cooperative tendency to individual and total outputs.²⁶ More empirical works are obviously needed to estimate the effects of these forces.

The paper suggests that resources should be allocated optimally between outside institutions and cooperative tendency, taking into consideration the dynamic interactions between them. For example, how many policemen are placed in the parks to punish littering is often affected by how many resources parents and teachers spend in teaching children not littering even when nobody is around, and vice versa. To rigorously address this problem, we need to endogenize both institutions and cooperative tendency, which is left for future research.

²⁶An anecdote may illuminate people’s view on this. Jack Welch, the former CEO of GE, recalls in his autobiography that expenditures (on meeting places and treadmills), aimed at encouraging informal socialization among employees across different departments, were strongly criticized as wasting money.

References

1. Andreoni, J. and R. Croson (2002), "Partners versus Strangers: Random Rematching in Public Goods Experiments," forthcoming in *Handbook of Experimental Economics Results*.
2. Arrow, K. (1972), "Gift and Exchanges," *Philosophy and Public Affairs*, I (1972), p343-62.
3. Bar-Gill, O., and C. Fershtman (2000), "The Limit of Public Policy: Endogenous Preferences," *working paper*, Aug. 2000.
4. Becker, G. (1996) *Accounting for Tastes*, Harvard University Press.
5. Bowles, S., and H. Gintis (1998) "The Determinants of Individual Earnings: Cognitive Skills, Personality, and Schooling," *working paper*.
6. Bowles, S., H. Gintis, and M. Osborne (2001) "The Determinants of Earnings: A Behavioral Approach," *Journal of Economics Literature*, Vol. XXXIX (Dec. 2001), pp 1137-1176.
7. Burlando, R., and J.D. Hey (1997), "Do Anglo-Saxons Free-ride More?" *Journal of Public Economics* 64 (1997) 41-60.
8. Cappelli, P. (1995), "Is the 'Skill Gap' Really About Attitudes?" *California Management Review* Vol. 37, No.4, Summer 1995.
9. Coleman, J.S., "Social Capital in the Creation of Human Capital," *American Journal of Sociology* 94 (1988): S95-S120.
10. Cook, Karen S. (2001), editor, *Trust in Society*, New York: Russell Sage Foundation, 2001.
11. Fehr, E. and S. Gächter (2000), "Fairness and Retaliation: The Economics of Reciprocity," *Journal of Economic Perspectives* 14, 159-182.
12. Frey, B.S. and F. Oberholtzer-Gee (1997), "The Cost of Price Incentives: an Empirical Analysis of Motivation Crowding Out," *The American Economic Review* 87, 746-755.
13. Glaeser, E.L., David Laibson, and Bruce Sacerdote (2000), "The Economic Approach to Social Capital," *NBER working paper* 7728.
14. Glaeser, E.L., David Laibson, J.A. Scheinkman, and C.L. Soutter (2000), "Measuring Trust," *Quarterly Journal of Economics*, Aug. 2000.
15. Heckman, James J. (1999) "Policies to Foster Human Capital," *NBER Working Paper* 7288, August 1999.
16. Huang, Fali (2002), "Estimations of Child Development Production Functions," *mimeo*, University of Pennsylvania.
17. Knack, S., and P. Keefer (1997), "Does Social Capital Have an Economic Payoff? A Cross-Country Investigation," *Quarterly Journal of Economics*, CXII (1997), 1251-1288.
18. Kreps, D.(1997), "Intrinsic Motivation and Extrinsic Incentives," *American Economic Review*, May 1997.

19. Kreps, D., P. Milgrom, J. Roberts, and R. Wilson (1982) "Rational Cooperation in the Finitely Repeated Prisoner's Dilemma." *Journal of Economic Theory*, vol. 27: p245-52.
20. La Porter, R., F. Lopez-de-Silanes, A. Shleifer, and R.W. Vishny, "Trust in Large Organizations," *American Economic Review*, May 1997.
21. Lucas, R.E. Jr. (1988), "On the Mechanics of Economic Development," *Journal of Monetary Economics* 22 (1988) 3-42.
22. Mailath, G. and A. Postlewaite (2001), "Social Assets," *CARESS working paper*, University of Pennsylvania.
23. OECD (2001), *The Well-being of Nations: the Role of Human and Social Capital*, Paris.
24. Palfrey, T.R. and J.E. Prisbrey, "Anomalous Behavior in Public Goods Experiments: How Much and Why?," *The American Economic Review*, 1997, 87/5, 829-846.
25. Putnam, R. D. (1993) (with R. Leonardi and R.Y. Nanetti), *Making Democracy Work*, Princeton, NJ: Princeton University Press, 1993.
26. Putnam, R. D. (1995), "Bowling Alone: America's Declining Social Capital," *Journal of Democracy*, Vol.6 (1995), pp. 65-78.
27. Rob, R., and P. Zemsky (2002), "Social Capital, Corporate Culture and the Incentive Intensity," *RAND Journal of Economics*, Vol. 33 No. 2, Summer 2002.
28. Rotemberg, J. J. (1994), "Human Relations in the Workplace." *Journal of Political Economy* 102 (August 1994): 684-718.
29. Temple, J. and P.A. Johnson (1998), "Social Capability and Economic Growth," *Quarterly Journal of Economics*, August,1998.
30. Tirole, J. (1996), "A Theory of Collective Reputations," *Review of Economic Studies* (1996) 63, 1-22.
31. Weimann, J. (1994), "Individual Behavior in a Free Riding Experiment," *Journal of Public Economics* 54, 185-200.
32. Welch, Jack (2001), *Straight from the Gut*, Warner Books, Inc., New York, NY.

Appendix:

• Proof for Claim 3.

Proof. We first prove that given the above specified belief system, non-selfish players cannot do better by deviation. (1) In period T following the history that only (C, C) has been played in all previous $T - 1$ periods, a player assigns probability π to his partner being non-selfish. In this case the last period game is the same as the above incomplete information one-shot game, where non-selfish players would play C according to Claim 2. (2) At any period $t < T$ following the history in which only (C, C) has been played, given his partner playing C according to the specified equilibrium strategy, a non-selfish player strictly prefers to play C since C dominates D due to $\alpha > d$. So non-selfish players will not deviate from the equilibrium strategy specified above by the assumption that they have $\alpha_i \geq \pi d + (1 - \pi)l$. If D is played by his partner, then a selfless player would still play his dominant strategy C , while reciprocal players would play D since (D, D) is a NE.

Next we prove that selfish players cannot do better by deviation if $\beta \geq \frac{d}{(g+d)\pi}$ given the specified belief system. At period T , playing D is selfish players' dominant strategy, so he will not deviate. Suppose he deviates at some period $t < T$ by playing D . But then his selfish type is revealed because of the belief system. According to the equilibrium strategies, if his partner is not selfless, (D, D) is played in the left periods after t ; only when his partner is selfless, (C, D) is played. Denote the proportion of selfless players by $\pi_S \equiv \Pr(\alpha_i > l)$. The deviation payoff for a selfish player from period t until T is

$$(g + d)\beta^{t-1} + (g + d)(\beta^t + \beta^{t+1} + \dots + \beta^{T-1})\pi_S.$$

By not deviating he can get

$$g\beta^{t-1} + g\beta^t + \dots + g\beta^{T-2} + (g + d)\pi\beta^{T-1} = g\beta^{t-1} \frac{1 - \beta^{T-t+1}}{1 - \beta} + d\pi\beta^{T-1}.$$

The non-deviation condition at period t for a selfish player is

$$\begin{aligned} (g + d)\beta^{t-1} + (g + d)(\beta^t + \beta^{t+1} + \dots + \beta^{T-1})\pi_S &< g\beta^{t-1} + g\beta^t + \dots + g\beta^{T-2} + (g + d)\pi\beta^{T-1} \\ \Rightarrow [g - (g + d)\pi_S] \frac{\beta(1 - \beta^{T-t-1})}{1 - \beta} + (g + d)(\pi - \pi_S)\beta^{T-t} &> d \end{aligned}$$

The LHS's partial derivation with respect to t is

$$\begin{aligned} \frac{\partial LHS}{\partial t} &= [g - (g + d)\pi_S] \frac{\beta^{T-t} \ln \beta}{1 - \beta} - (g + d)(\pi - \pi_S)\beta^{T-t} \ln \beta \\ &= \frac{-\beta^{T-t} \ln \beta}{1 - \beta} (g + d) \left[\pi - \frac{g}{g + d} - (\pi - \pi_S)\beta \right]. \end{aligned}$$

It is negative if

$$\beta \geq \left(\pi - \frac{g}{g + d} \right) / (\pi - \pi_S). \quad (7)$$

That is, if players are patient enough, they would wait until later to deviate, since deviation becomes more attractive as time goes by. In other words, if a selfish players do not want to deviate at period $T - 1$, then they will not deviate at any time earlier. Non-deviation at period $T - 1$ means

$$\begin{aligned} (g + d)\beta^{T-2} + (g + d)\beta^{T-1}\pi_S &< g\beta^{T-2} + (g + d)\pi\beta^{T-1} \\ \Rightarrow d &< (g + d)(\pi - \pi_S)\beta \\ \Rightarrow \beta &> \frac{d}{(g + d)(\pi - \pi_S)}. \end{aligned} \quad (8)$$

It is easy to check that condition (7) is implied by condition (8) because $\pi < 1$. So that the condition (8) guarantees that selfish players will not want to deviate at any time. To make sure that there is such β , $\frac{d}{(g+d)(\pi-\pi_S)}$ must be smaller than 1, which implies that $\pi - \pi_S > \frac{d}{g+d}$.

We have proved that the above specified strategy profile is sequentially rational w.r.t. the belief system. Now we show that the belief system is fully consistent given the strategy profile. In the first period, the probability that a player is non-selfish is equal to π since the match is random. In any period after the history that only (C, C) is played, the initial belief is still maintained because the two types of players cannot be distinguished from each other. If in some period (C, D) is observed, the player who plays D must be selfish since D is never a best response for a non-selfish player when his partner plays C . So the player who does not play D must update his belief about his partner's probability of being non-selfish from π to 0. While the probability of the player who plays C in (C, D) being non-selfish is still π because both types could do so according to the equilibrium strategy profile. ■

• **Proof for Proposition 5.**

Proof. Given any $\Pi \in [\Pi_0, \Pi_1]$, by implicit function theorem, we get from the equation $V_d(i^*, \Pi) = \beta[pG - (1-p)(\Pi d + (1-\Pi)l)] - c(D, i^*) = 0$ that

$$\begin{aligned} \frac{\partial B(\Pi)}{\partial d} &= \frac{\partial i^*(\Pi)}{\partial d} = -\frac{\partial V_d(i^*, \Pi)/\partial d}{\partial V_d(i^*, \Pi)/\partial i^*} = \frac{-(\beta(1-p) + c_D)\Pi}{-\partial V_d(i^*, \Pi)/\partial i^*} < 0, \\ \frac{\partial B(\Pi)}{\partial l} &= \frac{\partial i^*(\Pi)}{\partial l} = -\frac{\partial V_d(i^*, \Pi)/\partial l}{\partial V_d(i^*, \Pi)/\partial i^*} = \frac{-(\beta(1-p) + c_D)(1-\Pi)}{-\partial V_d(i^*, \Pi)/\partial i^*} < 0, \\ \frac{\partial B(\Pi)}{\partial \beta} &= \frac{\partial i^*(\Pi)}{\partial \beta} = -\frac{\partial V_d(i^*, \Pi)/\partial \beta}{\partial V_d(i^*, \Pi)/\partial i^*} = \frac{pG - (1-p)(\Pi d + (1-\Pi)l)}{-\partial V_d(i^*, \Pi)/\partial i^*} > 0, \\ \frac{\partial B(\Pi)}{\partial G} &= \frac{\partial i^*(\Pi)}{\partial G} = -\frac{\partial V_d(i^*, \Pi)/\partial G}{\partial V_d(i^*, \Pi)/\partial i^*} = \frac{\beta p}{-\partial V_d(i^*, \Pi)/\partial i^*} > 0. \end{aligned}$$

Again by implicit function theorem, we get from $\lim_{\Pi \rightarrow 0} V_d(i = \pi_0, \Pi) = \beta pG - \beta(1-p)l - c(l, \pi_0) = 0$ the following results

$$\begin{aligned} \frac{\partial \pi_0}{\partial d} &= 0, \\ \frac{\partial \pi_0}{\partial l} &= -\frac{-\beta(1-p) - c_l}{-c_i} < 0, \\ \frac{\partial \pi_0}{\partial \beta} &= -\frac{pG - (1-p)l}{-c_i} > 0, \\ \frac{\partial \pi_0}{\partial G} &= -\frac{\beta p}{-c_i} > 0. \end{aligned}$$

■

• **Proof for Proposition 8.**

Proof. (1) The Existence of Unique Solutions h_A^i and h_M^i .
The objective functions are

$$\begin{aligned} V_A^i(h) &= \beta(1-p)[\Pi g(h) - (1-\Pi)l(h)] + \beta pG(h) - c(h, \underline{\alpha}(h, \Pi), i), \\ V_M^i(h) &= \beta(1-p)\Pi[g(h) + d(h)] - c(h, 0, i). \end{aligned}$$

The FOC of V_M^i for an interior solution is,

$$[V_M^i(h)]' = \beta(1-p)\Pi[g'(h) + d'(h)] - c_h(h, 0, i) = 0 \quad (9)$$

Since $g''(h) \leq 0$, $d''(h) \leq 0$, and $c_{hh}(h, \alpha, i) > 0$, we know that $[V_M^i(h)]'$ is a decreasing function of h . If we assume that

$$\lim_{h \rightarrow 0} c_h(h, 0, i) = 0, \lim_{h \rightarrow 0} g'(h) > 0, \quad (A1)$$

we get $\lim_{h \rightarrow 0} V_M^i(h, 0) > 0$. So there is a unique solution $h_M^i = h_M(\Pi, p, \beta, T, i, k) \geq 0$ such that $V_M^i(h_M^i) = 0$, where k represents all other parameters.

The FOC of V_A^i for an interior solution is,

$$[V_A^i(h)]' = \beta(1-p)[\Pi g'(h) - (1-\Pi)l'(h)] + \beta p G'(h) - \frac{\partial c(h, \alpha(h, \Pi))}{\partial h} = 0. \quad (10)$$

The second derivative of value function $V_A^i(h)$ w.r.t. to h is

$$[V_A^i(h)]'' = \beta(1-p)[\Pi g''(h) - (1-\Pi)l''(h)] + \beta p G''(h) - \frac{\partial^2 c(h, \alpha(h, \Pi))}{\partial h^2},$$

where

$$\begin{aligned} \frac{\partial c(h, \alpha(h, \Pi))}{\partial h} &= c_h(h, \alpha(h, \Pi), i) + c_\alpha(h, \alpha(h, \Pi), i)\alpha_h(h, \Pi), \\ \frac{\partial^2 c(h, \alpha(h, \Pi), \Pi)}{\partial h^2} &= c_{hh} + (c_{h\alpha} + c_{\alpha h})\alpha_h(h, \Pi) + c_{\alpha\alpha}(\alpha_h(h, \Pi))^2 + c_\alpha\alpha_{hh}(h, \Pi). \end{aligned}$$

A sufficient condition to enable the second order condition to hold is

$$l''(h) = 0, \text{ and } \frac{\partial^2 c(h, \alpha(h, \Pi), \Pi)}{\partial h^2} \geq 0. \quad (A2)$$

To guarantee a non-negative solution, we have to assume that $[V_A^i(h=0)]' \geq 0$, which requires the boundary condition

$$\lim_{h \rightarrow 0} \frac{\partial c(h, \alpha(h, \Pi))}{\partial h} = 0, \lim_{h \rightarrow 0} [\beta p G'(h) + \beta(1-p)(\Pi(g'(h) + l'(h)) - l'(h))] > 0, \quad (A1')$$

Under these two conditions, we can get a unique solution $h_A^i = h_A(\Pi, p, \beta, T, i, k)$ such that $V_A^i(h_A^i) = 0$.

(2) Comparative Statics for h_A^i and h_M w.r.t. Π for any $i \in [0, 1]$.
By Implicit Function Theorem,

$$\begin{aligned} \frac{\partial h_M}{\partial \Pi} &= -\frac{\partial^2 [V_M^i(h)]}{\partial \Pi \partial h} / \frac{\partial^2 [V_M^i(h)]}{\partial h^2} = -\beta(1-p)[g'(h) + d'(h)] / \frac{\partial^2 [V_M^i(h)]}{\partial h^2} > 0. \\ \frac{\partial h_A^i}{\partial \Pi} &= -\frac{\partial^2 [V_A^i(h)]}{\partial \Pi \partial h} / \frac{\partial^2 [V_A^i(h)]}{\partial h^2} = -[\beta(1-p)(g'(h) + l'(h)) - \frac{\partial c(h, \alpha(h, \Pi), i)}{\partial h \partial \Pi}] / \frac{\partial^2 V_A^i(h)}{\partial h^2} > 0. \end{aligned}$$

if

$$\frac{\partial c(h, \alpha(h, \Pi), i)}{\partial h \partial \Pi} \leq 0, \quad (11)$$

where

$$\begin{aligned}\frac{\partial c(h, \alpha(h, \Pi), i)}{\partial h \partial \Pi} &= \frac{\partial [c_h(h, \alpha(h, \Pi), i) + c_\alpha(h, \alpha(h, \Pi), i) \alpha_h(h, \Pi)]}{\partial \Pi} \\ &= [c_{h\alpha}(h, \alpha(h, \Pi), i) + c_{\alpha\alpha}(h, \alpha(h, \Pi), i) \alpha_h(h, \Pi)](d(h) - l(h)) \\ &\quad + c_\alpha(h, \alpha(h, \Pi), i)(d'(h) - l'(h)).\end{aligned}$$

A sufficient condition for (11) to hold is

$$c_{h\alpha}(h, \alpha(h, \Pi), i) \geq 0, d'(h) \leq l'(h). \quad (\text{A3})$$

(3) Comparative Statics for h_A^i and h_M for any $i \in [0, 1]$ w.r.t. p

$$\begin{aligned}\frac{\partial h_M}{\partial p} &= -\frac{\partial^2 V_M^i(h)}{\partial p \partial h} / \frac{\partial^2 [V_M^i(h)]}{\partial h^2} = 0, \\ \frac{\partial h_A^i}{\partial p} &= -\frac{\partial^2 V_A^i(h)}{\partial p \partial h} / \frac{\partial^2 [V_A^i(h)]}{\partial h^2} = -\beta(G'(h) - \Pi g'(h) + (1 - \Pi)l'(h)) / \frac{\partial^2 V_A^i(h)}{\partial h^2} > 0.\end{aligned}$$

(4) Comparative Statics for h_A^i and h_M for any $i \in [0, 1]$ w.r.t. β

$$\begin{aligned}\frac{\partial h_M}{\partial \beta} &= -\frac{\partial^2 V_M^i(h)}{\partial \beta \partial h} / \frac{\partial^2 V_M^i(h)}{\partial h^2} = -\Pi(1 - p)[g'(h) + d'(h)] / \frac{\partial^2 V_M^i(h)}{\partial h^2} \geq 0, \\ \frac{\partial h_A^i}{\partial \beta} &= -\frac{\partial^2 V_A^i(h)}{\partial \beta \partial h} / \frac{\partial^2 V_A^i(h)}{\partial h^2} = -\{(1 - p)[\Pi g'(h) - (1 - \Pi)l'(h)] + G'(h)p\} / \frac{\partial^2 V_A^i(h)}{\partial h^2} \geq 0,\end{aligned}$$

since $\Pi(1 - p)(g'(h) + l'(h)) + pG'(h) - l'(h) > 0$ at h_A^i by condition (10). ■

• **Proof for Proposition 9.**

Proof. (1) The Relation Between h_M^i and h_M^j , h_A^i and h_A^j for any $i, j \in [0, 1]$

Since $[V_M^i(h)]' = 0$ by condition (9), we can use the Implicit Function Theorem and get

$$\begin{aligned}\frac{\partial h_M^i}{\partial i} &= -\frac{\partial [V_M^i(h)]'}{\partial i} / \frac{\partial [V_M^i(h)]'}{\partial h} \\ &= \frac{\partial c_h(h, 0, i)}{\partial i} / \frac{-\partial^2 V_M^i(h)}{\partial h^2} \stackrel{\geq}{\leq} 0, \\ \text{iff } \partial c_{hi}(h, 0, i) &\stackrel{\leq}{\geq} 0;\end{aligned}$$

Similarly from $[V_A^i(h)]' = 0$ we get

$$\begin{aligned}\frac{\partial h_A^i}{\partial i} &= -\frac{\partial [V_A^i(h)]'}{\partial i} / \frac{\partial [V_A^i(h)]'}{\partial h} \\ &= \frac{\partial^2 c(h, \alpha(h, \Pi), i)}{\partial h \partial i} / \frac{\partial^2 V_A^i(h)}{\partial h^2} \stackrel{\geq}{\leq} 0, \\ \text{iff } \frac{\partial^2 c(h, \alpha(h, \Pi), i)}{\partial h \partial i} &= c_{hi}(h, \alpha(h, \Pi), i) + c_{\alpha i}(h, \alpha(h, \Pi), i) \alpha_h(h, \Pi) \stackrel{\leq}{\geq} 0.\end{aligned}$$

Under the following assumption,

$$c_{hi}(h, 0, i) = 0, c_{hi}(h_A^i, \alpha(h_A^i, \Pi), i) = 0, c_{\alpha i}(h, \alpha(h), i) > 0, \quad (\text{A4})$$

we get that $h_M^i = h_M$ for any $i \in [0, 1]$, and $h_A^i > h_A^j$ for any $i < j$, where $i, j, \in [0, 1]$.

(2) The Relation Between h_A^i and h_M for any $i \in [0, 1]$

We know that $[V_A^i(h_A^i)]' = 0$ and $[V_A^i(h_A^i)]'' < 0$. If we can show that $[V_A^i(h_M)]' \geq 0$, then $h_A^i \geq h_M$ is proved. By condition (9), $[V_M^i(h)]' = 0$. It implies that

$$-\beta\Pi(1-p)[g'(h_M) + d'(h_M)] + c_h(h_M, 0, i) = 0.$$

Add this zero term to $[V_A^i(h_M)]'$, we get

$$\begin{aligned} [V_A^i(h_M)]' &= \underbrace{\beta p G'(h_M) - \beta(1-p)[\Pi d'(h_M) + (1-\Pi)l'(h_M)]}_{A(\Pi)} \\ &\quad - \underbrace{[\partial c(h_M, \alpha(h_M, \Pi), i) / \partial h_M - c_h(h_M, 0, i)]}_{B(i, \Pi)}. \\ &\equiv A(\Pi) - B(i, \Pi). \end{aligned} \tag{12}$$

The first term $A(\Pi)$ is the same for all players. The extra cost of investing in a positive α for player i , $B(i, \Pi)$, increases with players' index by assumption (??):

$$\frac{\partial B(i, \Pi)}{\partial i} = c_{\alpha i}(h_M, \alpha(h_M, \Pi), i) \alpha_h(h_M, \Pi) > 0.$$

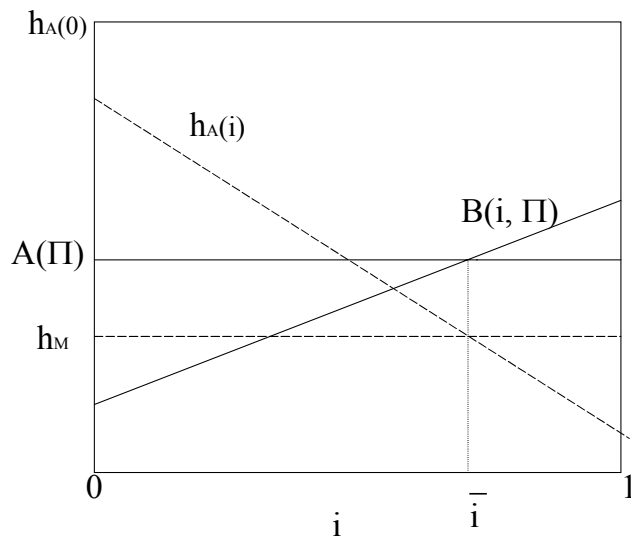
If the boundary condition $[V_A^0(h_M)]' \geq 0 \geq [V_A^1(h_M)]'$ holds given Π and p , i.e. if

$$B(1, \Pi) \geq A(\Pi) \geq B(0, \Pi), \tag{A5}$$

then there must exist a unique $\bar{i}(\Pi, p) \in [0, 1]$ such that

$$[V_A^{\bar{i}}(h_M)]' = A(\Pi) - B(\bar{i}, \Pi) = 0. \tag{13}$$

See figure below for illustration.



This condition means that for player \bar{i} , his optimal choice $h_{\bar{A}}^{\bar{i}}$ is equal to h_M . In other words, the benefit of being non-selfish is equal to the cost of investing in appropriate cooperative tendency so that his productive ability choice is not affected at all. Then for all $i \leq \bar{i}$, we have $[V_A^i(h_M)]' > 0 \iff h_A^i \geq h_M$; for all $i > \bar{i}$, $[V_A^i(h_M)]' < 0 \iff h_A^i < h_M$. If $A(\Pi) \geq B(1, \Pi)$, then $h_A^i \geq h_M$ for all i ; on the other hand, if $A(\Pi) < B(0, \Pi)$, the opposite is true.

Now we check the sign of $\frac{\partial \bar{i}}{\partial \Pi}$ based on equation (13).

$$\frac{\partial \bar{i}(\Pi, p)}{\partial \Pi} = -\frac{\partial V_A^i(h_M)/\partial h \partial \Pi}{\partial V_A^i(h_M)/\partial h \partial \bar{i}} = \frac{\partial V_A^i(h_M)/\partial h \partial \Pi}{c_{\alpha i}(h_M, \alpha(h_M, \Pi), i) \alpha_h(h_M, \Pi)} > 0$$

by condition (A4). Similarly, we get that

$$\frac{\partial \bar{i}(\Pi, p)}{\partial p} = -\frac{\partial V_A^i(h_M)/\partial h \partial p}{\partial V_A^i(h_M)/\partial h \partial \bar{i}} = \frac{\beta^2 G'(h_M)}{c_{\alpha i}(h_M, \alpha(h_M, \Pi), i) \alpha_h(h_M, \Pi)} > 0.$$

QED. ■

• **Proof for Lemma 2 (Human Capital Version).**

Proof. The derivative of $V_d(i, \Pi)$ with respect to i is

$$\begin{aligned} \frac{\partial V_d(i, \Pi)}{\partial i} &= \frac{\partial V_A^i(h_A^i) - \partial V_M^i(h_M)}{\partial i} \\ &= -[c_i(h_A^i, \alpha(h_A^i, \Pi), i) - c_i(h_M, 0, i)] < 0, \end{aligned}$$

by assumption (A4). The second equality holds according to the Envelope Theorem, since both $V_A^i(h_A^i)$ and $V_M^i(h_M^i)$ are maximized value functions. The effects of Π and i on abilities h_A^i and h_M^i have already been taken into consideration through the maximization process, and thus have no further power on the difference between the two optimal value functions.

Now we prove $\frac{\partial V_d(i, \Pi)}{\partial \Pi} > 0$. The derivative of $V_d(i, \Pi)$ with respect to Π is

$$\begin{aligned} \frac{\partial V_d(i, \Pi)}{\partial \Pi} &= \frac{\partial V_A^i(h_A^i) - \partial V_M^i(h_M)}{\partial \Pi} \\ &= \beta(1-p)[g(h_A^i) + l(h_A^i) - g(h_M) - d(h_M)] + c_{\alpha}(h_A^i, \alpha(h_A^i, \Pi), i)[l(h_A^i) - d(h_A^i)]. \end{aligned}$$

It is obvious that $\frac{\partial V_d(i, \Pi)}{\partial \Pi} > 0$ when $h_A^i \geq h_M$, which is true for low index players. If we can show that $\frac{\partial V_d(i, \Pi)}{\partial \Pi}$ reaches its infimum at the lowest index $i = 0$, then $\frac{\partial V_d(i, \Pi)}{\partial \Pi} > 0$ for all players. Indeed this is the case since $\frac{\partial^2 V_d(i, \Pi)}{\partial \Pi \partial i} > 0$.

$$\begin{aligned} \frac{\partial^2 V_d(i, \Pi)}{\partial \Pi \partial i} &\equiv \frac{\partial^2 V_d(i, \Pi)}{\partial i \partial \Pi} = \frac{-\partial[c_i(h_A^i, \alpha(h_A^i, \Pi), i) - c_i(h_M, 0, i)]}{\partial \Pi} = \underbrace{c_{i\alpha}(h_A^i, \alpha(h_A^i, \Pi), i)[l(h_A^i) - d(h_A^i) + \alpha_h \frac{\partial h_A^i}{\partial \Pi}] +}_{> 0} \\ &\quad \text{because } c_{i\alpha}(h, \alpha(h, \Pi), i) > 0, \frac{\partial h_A^i}{\partial \Pi} \geq 0 \\ &\underbrace{c_{ih}(h_A^i, \alpha(h_A^i, \Pi), i) \frac{\partial h_A^i}{\partial \Pi} - c_{ih}(h_M, 0, i) \frac{\partial h_M}{\partial \Pi}}_{= 0} \\ &\quad \text{because } c_{ih}(h, 0, i) = 0 \end{aligned}$$

The intuition behind $\frac{\partial^2 V_d(i, \Pi)}{\partial \Pi \partial i} > 0$ is that the high cost players get more benefit from the reduced $\alpha(h_A^i, \Pi)$ due to a higher Π . ■