Social Trust, Cooperation, and Human Capital[*]

Fali Huang
Department of Economics
Singapore Management University

December 11, 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. For utility-maximizing players, cooperation arises in one-period prisoner's dilemmas if and only if there is social trust. The amount of social trust in a given game is determined by the distribution of players' cooperative tendency. Cooperative tendency is in essence a component of human capital distinct from cognitive ability. Its investment, however, is typically not efficient because the social returns are always strictly larger than individual returns. This positive investment externality leads to multiple equilibria in social trust formation, but a unique stable equilibrium may also exist. The different effects of legal institutions, information structure and education programs on social trust are also investigated. (*JEL* Z13, J24)

1 Introduction

The importance of social trust in economy was suggested long ago by Arrow (1972, p.357): "Virtually every commercial transaction has within itself an element of trust, certainly any transaction conducted over a period of time. It can be plausibly argued that much of the economic backwardness in the world can be explained by the lack of mutual confidence." In recent years social trust, as an important form of social capital, has attracted the attention

[*] 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.

of many economists as well as other social scientists.[1] For example, several empirical works show that the average trusting level in a society is significantly associated with economic growth (Knack and Keefer, 1997), and has large positive effects on the performance of various organizations (La Porta et al., 1997).

The formal (economic) analysis of social trust, however, is lagging behind and answers to many basic questions about social trust are still elusive. For example, what is the relationship between trustworthiness, trust, and social trust? What is the unique role of social trust in promoting cooperation compared with other forces such as reputation? How does social trust vary across games and players? How can one account for the discrepancies or even contradictions among different empirical measures of social trust? How is social trust formed in society? What are the roles of education systems, mass media, social networks, and legal institutions in the formation of social trust? Is the social trust level in a society efficient? Is it history dependent? How is it related to human capital?

Motivated by these questions, this paper formalizes several trust-related concepts, and studies the formation of social trust in a game theoretic setting. Note that the concept of *trust* is vacuous without discrepancies 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 as trust among strangers instead of acquaintances involved in repeated interactions, or the residual trust unexplained by these arrangements (Hardin 2002). So a one-period prisoner's dilemma seems to be an ideal context to formalize social trust.

People in general differ in their predisposition to cooperate, i.e., they have different cooperative tendencies (Palfrey and Prisbrey 1997). In a one-period prisoner's dilemma, players with sufficiently high (low) cooperative tendencies will cooperate (defect) regardless of their partner's action, while those in the middle behave in a reciprocal way. A player's *trustworthiness* in a game can be defined as the probability that he will cooperate in it.

[1] 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).

So the same player may exhibit different levels of trustworthiness across games; those with higher cooperative tendencies are more trustworthy. How much *trust* we have in a player is equal to his trustworthiness. The *social trust* among a group of players is the expected trustworthiness of a typical member, which is determined by the distribution of cooperative tendency among players besides relevant game features. Its amount may vary across games and players, which explains why there are discrepancies in social trust measures based on surveys and experiments (Glaeser et al., 2000).

The effects of social trust in generating cooperation can be amplified by repeated interactions among players. In finitely repeated games, social trust is a *necessary* condition for the existence of reputation effects among utility-maximizing players. Indeed, the crucial types typically *assumed* in the reputation literature, for example, the tit-for-tat type in Kreps et al. (1982), the honest one in Tirole (1996) and Dixit (2003), and the reciprocal one in Fehr and Gachter (2000), are simply special cases of trustworthy players. In this sense, social trust among players serves as a solid base for reputations to build on.

As an idiosyncratic feature of individuals, a cooperative tendency is often costly to cultivate and may generate returns in the future. It is, in essence, a component of human capital that is distinct from cognitive skills. The cooperative tendency enables a player to cooperate more and get higher returns for his cognitive skills, while it also competes with cognitive skills for resources at human capital investment stage. For players with low investing cost in cooperative tendency, being trustworthy would increase their cognitive skills than remaining selfishness. Investment in a cooperative tendency, however, is not efficient because social returns are always strictly larger than individual returns. This positive externality implies that equilibrium social trust levels are typically not optimal, and there exist multiple equilibria. In the benchmark case where individual returns 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, there is only one stable equilibrium when individual returns are diverse.

When investing costs are reduced by education systems, or when the observability of cooperative tendency is improved by information structure (including mass media and social networks), both social trust and average cooperative tendency increase. In contrast, when legal institutions and monitoring reduce the benefits of defecting, social trust levels in relevant situations are increased, but individuals' internal discipline may be crowded out.

Glaeser, Laibson, and Sacerdote (2000) are among the first economists that formally studied social capital formation. The current paper differs from theirs in a couple of important aspects. First, they model individual investment decisions as isolated optimization problems. In contrast, we adopt a game theoretic setting and an equilibrium approach to study social trust formation, which is consistent with Coleman's (1988) insight that social capital "exists in the *relations* among persons." Indeed, strong externalities among players may be the defining feature of *social* capital. Second, we focus on social trust alone rather than bundle different forms of social capital together as a homogenous subject. Social capital is arguably an umbrella concept whose many manifestations differ substantially 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. It seems that social trust best captures the essence of social capital and the rigorous study of social trust may be the key to eventually understand other social capital forms.

Rob and Zemsky (2002) study the effects of incentive structures in a firm on social trust among employees. They find that different corporate cultures can be caused purely by ex ante differences in employees' cooperative tendencies. The current paper complements their work in that it focuses on individual optimal choices and endogenizes cooperative tendencies and social trust in a society. In this aspect it is similar to the work of Frank (1987) that studies the endogenous choice of being honest by selfish players. There are several differences, however. For example, the cooperative tendency as a continuous variable is more general than the binary trait of honest and selfish; its relationship with social trust and human capital, which is the focus of this paper, is not mentioned at all by Frank.

From the perspective of the human capital literature (Heckman, 2000; Bowles et al., 2001), this paper provides new evidence that non-cognitive skills, including incentive-enhancing preferences, are important determinants of individual earnings. It also detects large positive externalities in cooperative tendency investment, which may account for the shortage of appropriate working attitudes among employees in many firms (Cappelli, 1995).

The paper is organized as follows. In the next section cooperative tendency and social trust are formally defined and investigated in various games. A simple social trust formation model is developed in section three, where players' cooperative tendencies are chosen (by their parents) to maximize their life time utility. 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. g, l, and d are payoffs or material outputs for players, where g and l represent respectively *g*ain and *l*oss of cooperative behavior, while d is for extra gain from *d*efecting. We make the following two assumptions

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

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

which are quite standard in relevant literature (Kreps et al 1982, Rotemberg 1994, Bar-Gill and Fershtman 2000). The first assumption generates reciprocal behaviors.[2] Together with the second one, it guarantees that cooperative behaviors always increase total payoffs.

[2] Note that d and l represent a player's marginal costs of cooperating when his partner cooperates and defects, respectively. A similar analysis is by Dixit (2003).

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; $m_i(A_i, A_j) \in \{g, g + d, -l, 0\}$ is the game-specific payoff for player i; $\chi_D(A_i)$ is an index function such that

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

$\alpha_i \in R^+$ is the amount of disutility player i incurs when defecting, which measures player i's taste for cooperation, or *cooperative tendency.* It is an internal discipline against defecting that may enable players to cooperate in situations where cooperation is otherwise not chosen.[3] Players have heterogenous cooperative tendencies such that $\alpha_i \sim F(\cdot)$, where $F(\cdot)$ is a cumulative distribution function.

Each player's payoff from the game is thus composed of two parts: $m_i(A_i, A_j)$ is game-specific and constant across players, while $\alpha_i \chi_D(A_i)$ is player-specific and stable across games.[4] These two components are explicitly modeled here to highlight two distinct ways of inducing cooperation in prisoner's dilemmas (James 2002). The conventional way changes game-specific payoffs by embedding a dilemma in a bigger game. For example, appropriate rewards and punishments associated with repeated interactions can transform a stand-alone prisoner's dilemma to a new game where cooperation becomes a Nash equilibrium. A second way modifies player-specific payoffs in that people may care about things other than narrow selfish payoffs (e.g. Frank 1987, Kandel and Lazear 1992, Rotemberg 1994, Bar-Gill and

[3] This utility function is motivated by the experimental finding that warm-glow effects are highly significant in inducing cooperation in public good games (Palfrey and Prisbrey 1997). It does not make any difference in this section if we model cooperative tendency as an intrinsic *benefit* to cooperation. But it may cause unnecessary technical complication in the next section when individuals have to decide how much cooperative tendency they would invest, since they may want to invest in cooperative tendency *per se* to simply feel good, rather than as a valuable asset enabling them to cooperate in a productive way.

[4] Accordingly, the payoffs associated with the same actions in the above prisoner's dilemma differ across players. To avoid confusion and be consistent with the standard usage in literature, we call a game a prisoner's dilemma if it is so for players with zero cooperative tendency.

Fershtman 2000). As long as these 'special' preferences can be observed and/or intentionally cultivated in reality, this way would yields insightful and refutable results as well.

2.2 Trustworthiness, Trust, Social Trust

Fix a one-period prisoner's dilemma game. Players with different cooperative tendencies can be categorized into three behavioral types: the *selfish* if he always defects, the *selfless* if he always cooperates, and the *reciprocal* if he makes in-kind responses to his partner's action.[5] The latter two types are also called *cooperative* or *non-selfish*.

The *trustworthiness* of a player in the game is the probability that he would cooperate in it. Selfish (selfless) players have zero (full) trustworthiness since they never (always) cooperate; the trustworthiness of a reciprocal player is zero (one) if matched with a selfish (cooperative) partner. How much *trust* a player has in his partner is equal to the latter's trustworthiness. So all players have no (full) trust in a selfish (selfless) partner. A cooperative player will trust a reciprocal partner, but a selfish one will not. *Social trust* in a group is equal to the expected trustworthiness of a typical member, which is determined by the distribution of cooperative tendency $F(\cdot)$ and specific game features such as defecting benefits and information structure. It can be characterized by the proportions of reciprocal and selfless players.

2.3 Social Trust in Various Games

This section studies the levels of social trust and its effects on outputs in various games. Specifically it proves the following proposition.

Proposition 1 *Fix the distribution of cooperative tendency in population. In one-period prisoner's dilemmas cooperation arises if and only if there is social trust among players. In finitely repeated games, social trust is a necessary condition for reputation effects. Social trust varies across games, decreasing in defecting benefits d and l. The total output strictly increases with social trust.*

[5] Many 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 Gachter 2000).

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	$(g+d-\alpha_A, -l)$	$(-\alpha_A, -\alpha_M)$

Proposition 2 *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*) When Mike plays C, Ann will play C iff $g \geq g + d - \alpha_A \Leftrightarrow \alpha_A \geq d$ holds. When Mike plays D, Ann will play C iff $\alpha_A \geq l$ holds. So Ann's best response is: always defect when $\alpha_A < d$; always cooperate when $\alpha_A \geq l$; reciprocate otherwise. Since the game is symmetric, Mike has the same best response function. *ii*) Since selfless players always cooperate and reciprocal players always reciprocate with cooperative behaviors, (C, C) is a Nash equilibrium when both players are non-selfish. Since a selfish player never plays C, (C, C) can not be a Nash equilibrium if at least one players is selfish. ■

Let π_{RC} denote the proportion of the reciprocal type under complete information and π_{SC} the selfless type. Then $\pi_{RC} = \Pr(d \leq \alpha_i < l) = F(l) - F(d)$, $\pi_{SC} = \Pr(\alpha_i \geq l) = 1 - F(l)$ by proposition 2. When players are randomly matched with each other, social trust is π_{SC} from a selfish players' perspective, and $(\pi_{RC} + \pi_{SC})$ for non-selfish players. It is obvious that π_{SC} decreases with l and $(\pi_{RC} + \pi_{SC})$ decreases with d. The expected outputs are $\pi_{SC}(g + d)$, $(\pi_{RC} + \pi_{SC})g$ and $(\pi_{RC} + \pi_{SC})(g + l) - l$ respectively for selfish, reciprocal,[6] and selfless players, all strictly increasing with social trust π_{SC} or $(\pi_{RC} + \pi_{SC})$.

[6] Note 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.

An alternative matching system is assortative matching where selfish players match with each other and cooperative ones match among themselves. In this case, social trust is zero among selfish players and one among cooperative players. The amount of social trust in the whole group is $(\pi_{RC} + \pi_{SC})$. Since cooperative players each produce output g and selfish ones zero, the total output is $g(\pi_{RC} + \pi_{SC})$, again strictly increasing in social trust.

2.3.2 One-period Incomplete Information Game

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

Proposition 3 *In the one-period incomplete information game, 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 the probability of a player matching with a cooperative partner is believed to be π. By playing C, a player i gets g if her partner is cooperative, $-l$ if her partner defects. So her expected payoff of playing C is $\pi g - (1-\pi)l$. Similarly we get that her expected utility of playing D is $\pi(g + d - \alpha_i) - (1-\pi)\alpha_i$. She will play C iff $\pi g - (1-\pi)l \ \geq \pi(g + d - \alpha_i) - (1-\pi)\alpha_i$ or $\alpha_i \geq \pi d + (1-\pi)l$ holds.

For belief π to be 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 and increasing in π on the closed interval $[0,1]$ because $\frac{\partial RHS}{\partial \pi} = (l-d)DF \geq 0$. We also have $RHS(\pi = 0) = 1 - F(l) \geq 0$ and $RHS(\pi = 1) = 1 - F(d) \leq 1$, which implies that the slope at π is smaller than one, i.e. $(l-d)DF < 1$. So π is uniquely determined in the interval $[1 - F(l), 1 - F(d)] \subseteq [0,1]$, and it is stable. By the Implicit Function Theorem we have $\frac{\partial \pi}{\partial d} = -\frac{\pi dF}{1-(l-d)dF} < 0$ and $\frac{\partial \pi}{\partial l} = -\frac{(1-\pi)dF}{1-(l-d)dF} < 0$. ■

Let $\underline{\alpha}$ denote the minimum cooperative tendency for a player to become cooperative, then $\underline{\alpha} \equiv \pi d + (1-\pi)l$. Proposition 3 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 is now characterized by $\pi = \pi_{SI} + \pi_{RI} = 1 - F(\underline{\alpha})$, the proportion of cooperative players.[7] The expected output of a selfish player $\pi(g+d)$ is higher than that of a cooperative player $\pi g - (1-\pi)l$. Both, however, strictly increase with social trust π. And cooperative players get higher marginal benefit from social trust.

2.3.3 T-period Incomplete Information Game

In one-shot games social trust improves output by enabling non-selfish players to cooperate in situations that would otherwise be a prisoner's dilemma. In repeated games social trust can elicit cooperative behavior even from selfish players through reputation effects. A sequential equilibrium in a finite T-period game with incomplete information is characterized below to illustrate interactions between social trust and repeated games. 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.

Proposition 4 *In this T-period game, the following strategy profile and belief system is a sequential equilibrium if $\beta\pi_{RI} \geq \frac{d}{(g+d)}$. 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 being 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.*

[7] 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 $\pi_{SC} < \pi < \pi_{SC} + \pi_{RC}$. So compared with complete information, social trust under incomplete information is lower for cooperative players, but higher for selfish players.

Proof. In the appendix. ■

In this equilibrium all players cooperate until the last period when they behave according to proposition 3. When we look at each period *in isolation*, it seems that repeated interactions promote 'trust' among players. But the expected trustworthiness of players (and thus the social trust) is always π on the equilibrium path. This discrepancy arises because there are two different sources of cooperation. One is trust based on the players' cooperative tendencies or goodwill. The other is the reputation effect, i.e., the scheme of rewards and punishments contingent on past behaviors, that makes cooperation appealing to a player's narrow selfish interests. So the true motivation for selfish players to cooperate is reputation concerns, not that they have become more trustworthy. Note that without enough (punishment from) reciprocal players, the reputation effect vanishes and selfish players will not cooperate anymore.

In repeated games, these two sources of cooperation are often mixed together. In one-shot games, however, cooperation arises only when there is trust among players. This is why trust should be defined and measured in one-shot games, which helps disentangle different sources of cooperation. For example, when repeated interactions are stopped unexpectedly, we can confidently predict that selfish players will not cooperate anymore, but non-selfish ones will continue to cooperate. When there is no social trust, however, nobody would cooperate in finitely repeated games. Since many institutions such as social networks and norms involve repeated interactions, our analysis may shed light on how they interact with social trust in promoting cooperation.

Another observation is that reciprocal players act similarly as tit-for-tat players in Kreps et al.(1982); and selfless players act as the honest type in Frank (1987), Tirole (1997) and Dixit (2003). In other words, our formulation of cooperative tendency naturally generate these irrational behaviors for rational players. This suggests an innate link between social trust and economic governance since the latter is affected by the proportion of honest type (Dixit 2003).

2.4 Empirical Measures of Social Trust

Our formalization of social trust is quite useful in clarifying the relationship between various empirical measures of social trust, especially when discrepancies among survey-based and experiment-based measures are quite common (Glaeser et al. 2000b, Burlando and Hey 1997, Weimann 1994).

The widely used trust indicator $TRUST$ is measured as below. The World Values Surveys ask people the following trust question: "Generally speaking, would you say that most people can be trusted, or that you can't be too careful in dealing with people?" $TRUST_C$ is equal to the percentage of respondents in country C replying "most people can be trusted." It exactly measures the amount of social trust in a country under some assumptions described below. 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 country C the representative dilemma is γ_C and the proportion of cooperative players in equilibrium is π_C. Players who have met a partner that can be trusted would agree that "most people can be trusted." Then $TRUST_C = \pi_C$ holds since exactly π_C proportion of players meet a trustworthy partner.

In a public goods experiment γ_P, suppose π_P proportion of players cooperate. If the distribution of cooperative tendency among the subjects is a random sample drawn from the whole population, and γ_P is the same as the representative dilemma γ_C in country C, then π_P is an unbiased estimate of $TRUST_C$. If any of these assumptions are violated, discrepancies among different measures of social trust would inevitably arise.

3 Social Trust Formation

Suppose parents are able to teach children to become more cooperative by acting as role models and choosing appropriate home and school inputs. How many of them would choose to do so in equilibrium?[8] Is the equilibrium social trust level optimal? If not, how can we

[8] General Social Surveys from 1986 to 1998 in the U.S. show that parents try to invest certain desirable traits in children. For example, 77.2% of parents consider "help others when they need help" one of the

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 his parents) to maximize his life-time utility, taking as given the expected proportion of cooperative players $\Pi \in [0,1]$ in the population.[9] Investing in cooperative tendency incurs positive cost. For example, parents have to repeatedly make effort in teaching children to share toys and be considerate. This task is easier when parents are more skillful and the child is more obedient. Let the cost function be $c(\alpha, i)$, where $c(0,i) = 0$, $c_\alpha > 0$, $c_i > 0$, $c_{\alpha\alpha} \geq 0$, $c_{\alpha i} \geq 0$. That is, the cost increases with player index i and is convex in the cooperative tendency α.

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, however, players' cooperative tendencies are publicly observed[10] and they are free to choose 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, if a player ever invests, his cooperative tendency will be equal to D which just enables him to cooperate in the complete information game (G, D, L).

Lemma 1 *$\alpha_i = D$ iff $\alpha_i > 0$ for any $i \in [0,1]$.*

Proof. Players with $\alpha \geq D$ cooperate in both games and get utility $pG + (1-p)[\Pi g - (1-\Pi)l]$. Since this payoff does not depend on α and investing in α is costly, it is optimal to choose the lowest possible level D. Players with $\alpha < \Pi d + (1-\Pi)l$ always defect and get

three most important traits that their children should learn.

[9] We assume a person's cooperative tendency is fixed before adulthood, which is consistent with casual observation. Individuals' trusting levels and trustworthiness may nonetheless change in accordance to different games, players and updated information (Alesina and La Ferrara 2002).

[10] Mailath et al (2003) show that it is impossible to maintain a permanent reputation for playing a strategy that does not play an equilibrium of the game with no uncertainty about types. In other words, a player's true type would ultimately be revealed by his actions. See also Frank (1987) for more reasons why p is positive and an elaborate treatment of information structure.

$(1-p)\Pi(g+d)-p\alpha$. Their payoffs are maximized when $\alpha = 0$. Any cooperative tendency in the middle range makes players worse off than otherwise. Players with such an α cooperate in game (g, d, l) but not in (G, D, L), getting payoff $(1-p)[\Pi g - (1-\Pi)l] - p\alpha$ which is worse than the cases of $\alpha = 0$ and $\alpha = D$. ■

We assume $D = \Pi d + (1-\Pi)l$ without much loss of generality. 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. By definition

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

Players will choose to invest if and only if $V_d \geq 0$.

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

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

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

3.2 Positive Externality

Suppose the social planner's objective function is to maximize the sum of all players' life-time 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 the highest index among them. Then $\pi = \Pr(i \leq i^S) = i^S$, since i is uniformly distributed on the interval $[0, 1]$.

Proposition 5 *The social returns of investment in the cooperative tendency are strictly larger than individual returns.*

Proof. The first derivative of social welfare $V(\pi)$ with respect to π is

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

The social return of player i^S investing in cooperative tendency is composed of two parts: the individual return i^S gets, and externalities of his investment on all others. The externalities are positive because all players benefit from an increase in social trust: $\frac{\partial V(D,i)}{\partial \pi} = \beta(1 - p)(g + l) + c_D(l - d) > 0$, $\frac{\partial V(0,i)}{\partial \pi} = \beta(1 - p)(g + d) > 0$. So the social return for any player being non-selfish is always strictly larger than his individual return. ■

This proposition implies that individual investment in cooperative tendency is generally not efficient, which may explain why there is shortage of appropriate working habits and attitudes in many firm (Cappelli 1995). Another implication is that equilibrium social trust is always strictly lower than social optimal level, except when there is already full trust in equilibrium.[11]

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 ($\Pi = e, \pi = e$), $e \in [0, 1]$, where π is the actual proportion of non-selfish players.[12] And 'no social trust'

[11] Though it is probable that full trust is the social optimal solution, this may not always be the case. For example, if some players have extremely high investing costs, say higher than positive externalities received by all others, then it is better that they remain selfish.

[12] Given 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 $\pi = \Pi$.

equilibrium ($\Pi = 0, \pi = 0$) always exists since there is no gain from being the only cooperative player. We partition the parameter space into four cases and characterize the corresponding equilibria. We also check whether these NEs are stable to small perturbations of Π.[13]

3.3.1 The Benchmark: Medium Cost Case

In this case the net returns of investing in cooperative tendency are quite similar across players, not too high or too low. Specifically it is characterized by the following conditions

$$V_d(0, \Pi_0) = 0, \tag{3}$$

$$V_d(1, \Pi_1) = 0. \tag{4}$$

That is, there exist $\Pi_0, \Pi_1 \in (0,1)$ such that no players want to invest when $\Pi \leq \Pi_0$, and all choose to invest when $\Pi \geq \Pi_1$. It is easy to prove $\Pi_0 < \Pi_1$. By Lemma 2 the inequality $V_d(1, \Pi_0) < V_d(0, \Pi_0)$ holds. But since $V_d(0, \Pi_0) = 0 = V_d(1, \Pi_1)$, we have $V_d(1, \Pi_0) < V_d(1, \Pi_1)$, which implies $\Pi_0 < \Pi_1$ by Lemma 2.

Lemma 3 *There is a unique solution* $i^*(\Pi)$ *to* $V_d(i(\Pi), \Pi) = 0$ *for any* $\Pi \in [\Pi_0, \Pi_1]$. $i^*(\Pi)$ *strictly increases in* Π.

Proof. By Lemma 2, $V_d(i, \Pi)$ is continuous and strictly decreasing in $i \in [0,1]$ for any Π. We also 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 $V_d(i^*(\Pi), \Pi) = 0$. $i^*(\Pi)$ strictly increases in Π since

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

by Lemma 2. ■

This lemma implies that for any $\Pi \in [\Pi_0, \Pi_1]$ all players with lower index than $i^*(\Pi)$ will choose to become cooperative while others will not. So the proportion of cooperative

[13] NEs can be considered as steady states in a dynamic process of countable infinite generations where the expected social trust in every following generation is equal to the realized one in its immediate predecessor. That is, $\Pi_{N+1} = \pi_N, \forall N = 1, 2, ...$ where the initial one $\Pi_{N=1}$ is assumed exogenously given.

players is $\pi = B(\Pi) = \Pr(i \leq i^*(\Pi) = i^*(\Pi)$ since i is uniformly distributed on $[0,1]$. It is trivial to show that $B(\Pi) = 0$ for any $\Pi \in [0, \Pi_0]$, and $B(\Pi) = 1$ for any $\Pi \in [\Pi_1, 1]$. Thus the best response function is

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

Since $B(\Pi)$ is continuous, strictly increases in Π on $[\Pi_0, \Pi_1]$, plus $B(\Pi_0) = 0$ and $B(\Pi_1) = 1$, there must exist at least one fixed point $\Pi^* \in [\Pi_0, \Pi_1]$ such that $i^*(\Pi^*) = \Pi^*$. 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.[14] It is not stable since the slope of $B(\Pi)$ is bigger than one when crossing the 45^0 line. It is easy to check that $(\Pi = 0, \pi = 0)$ and $(\Pi = 1, \pi = 1)$ are the other two NEs. Thus we have proved the following proposition.

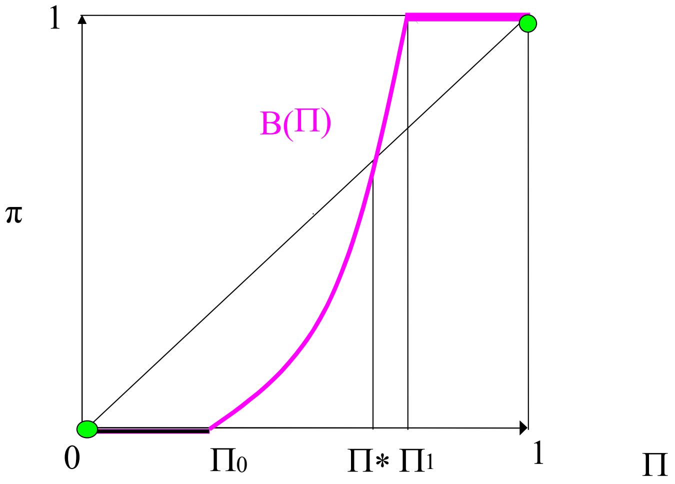


Figure 1: Medium Cost Case

Proposition 6 *Under conditions (3) and (4) there are three NEs:* $(0,0)$, (Π^*, Π^*), *and*

[14] The exact number of fixed points on $[\Pi_0, \Pi_1]$ depends on the curvature of $i^*(\Pi)$, which is difficult to pindown in general. Note that $\partial^2 i^*(\Pi)/\partial \Pi^2 = \beta(l-d)^2[(1-p+c_\alpha)c_{i\alpha} - c_{\alpha\alpha}c_i]/c_i^2$ is positive if $c_{\alpha\alpha} = 0$ or $c(D,i) = D^2 + Di$. When the best response function is linear, it must be $i^*(\Pi) = a\Pi - b$, where $a = \frac{1}{\Pi_1 - \Pi_0}$ and $b = \frac{\Pi_0}{\Pi_1 - \Pi_0}$ by conditions (3) and (4). Note that the slope a is bigger than 1. And $\Pi_l^* = \frac{\Pi_0}{1 + \Pi_0 - \Pi_1}$ solves $i^*(\Pi_l^*) = \Pi_l^*$.

$(1,1)$, *where* $\Pi^* \in [\Pi_0, \Pi_1] \subset (0,1)$. *Among them* $(0,0)$ *and* $(1,1)$ *are stable.*

The benchmark case is illustrated by figure 1. The interior NE (Π^*, Π^*) is unstable, happening only when the initial belief is exactly Π^*. If the initial belief is $\frac{\varepsilon}{2}$ lower than Π^*, this economy will ultimately fall into '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 $(\Pi = 1, \pi = 1)$. So a negligible ε difference in initial beliefs may lead to two polar stable equilibria (Putnam 1993). The intuition is that, when enough people (over the threshold Π^*) invest in cooperative tendency, the associated positive externalities outweigh idiosyncratic cost differences and make net returns positive for everybody, 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 have costs so low that they would invest in cooperative tendency no matter how few players are expected to do so. On the other hand, there are players whose costs are so high that they would not invest even everybody else does so. This case is characterized by the following conditions.

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

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

Let π_0 be defined by $lim_{\Pi \to 0^+} V_d(i^*(\Pi), 0) = 0$ and $\pi_0 \equiv lim_{\Pi \to 0^+} i^*(\Pi)$; and π_1 by $V_d(i^*(1), 1) = 0$ and $\pi_1 \equiv i^*(1)$. Conditions (5) and (6) are equivalent to $\pi_0 > 0$ and $\pi_1 < 1$, respectively. That is, as long as the expected social trust is positive, there are at least 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. The idiosyncratic differences in investment cost now outweigh the externalities, making players less affected by other people's choices.

Proposition 7 *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.*

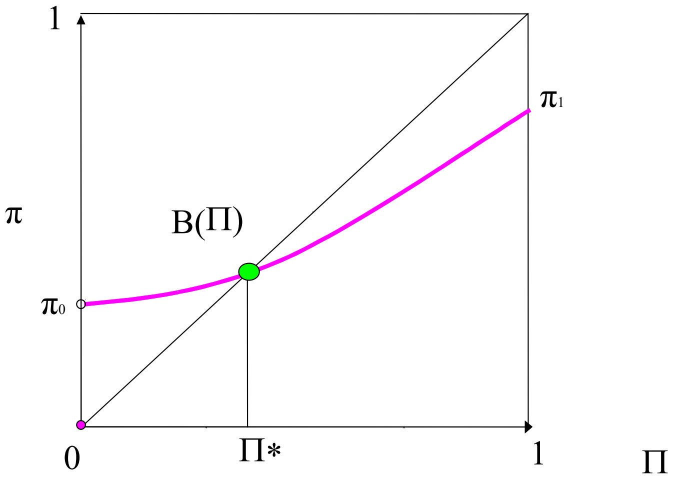


Figure 2: Diverse Cost Case

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

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

Let $\Pi^* \in (0,1)$ denote the solution to $i^*(\Pi^*) = \Pi^*$, then $\Pi^* \in (\pi_0, \pi_1)$ is unique because $B(\Pi \to 0) = \pi_0$, $B(\Pi = 1) = \pi_1$, and $B(\Pi)$ strictly increases in $\Pi \in (0,1]$. It is stable since the slope of $B(\Pi)$ is smaller than one when crossing the 45^0 line.[15] ■

This proposition shows that the interior NE (Π^*, Π^*) is the only focal point of the history, which is determined by fundamental forces such as information and cost structures and thus immune to random events. See figure 2 for illustration. The following comparative statics suggest that long-run social trust increases in p, β and G (representing expected returns of cooperative tendency), and decreases with defecting benefits d and l. These results are also true for any interior stable equilibrium in other cases.

Proposition 8 $\partial\Pi^*/\partial p > 0$, $\partial\Pi^*/\partial\beta > 0$, $\partial\Pi^*/\partial G > 0$; $\partial\Pi^*/\partial d < 0$, $\partial\Pi^*/\partial l < 0$.

Proof. To prove $\partial\Pi^*/\partial p > 0$, we show that p shifts up the best response function $B(\Pi)$ for each $\Pi \in (0,1]$ and increases π_0. Accordingly, the intersection of $B(\Pi)$ with the 45^0 line,

[15] The linear best response function is $i^*(\Pi) = c\Pi + d$, where $d = \pi_0, c + d = \pi_1$. Let $i^*(\Pi_l^*) = \Pi_l^*$, we get $\Pi_l^* = \frac{\pi_0}{1+\pi_0-\pi_1}$.

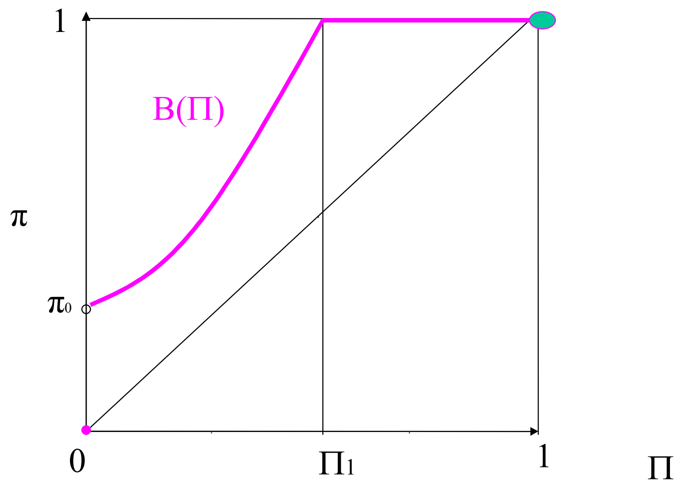


Figure 3: Low Cost Case

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 $lim_{\Pi \to 0} V_d(i = \pi_0, \Pi) = \beta pG - \beta(1-p)l - c(l, \pi_0) = 0$. 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 similarly and thus are relegated to the appendix. ■

3.3.3 Low Cost and High Cost Cases

In the low cost case, even the highest indexed players invest in cooperative tendency when they believe enough people are doing so. It is characterized by conditions (4) and (5) where $\Pi_1 \in [0, 1]$. Here full trust *NE* $(1, 1)$ always exists and is stable. It is either the only equilibrium, or there exist two other *NE*s at interior points where the one with lower social trust is stable. See figure 3. The high cost case is defined by conditions (3) and (6) where $\Pi_0 \in (0, 1]$. The no-trust *NE* $(0, 0)$ is stable. It may be unique, otherwise two interior *NE*s also exist where the one with higher social trust is stable. See figure 4 for illustration. The proof is omitted since it is similar to the first two cases.

Proposition 9 *i) Multiple equilibrium are possible in all cases. However, among all NEs*

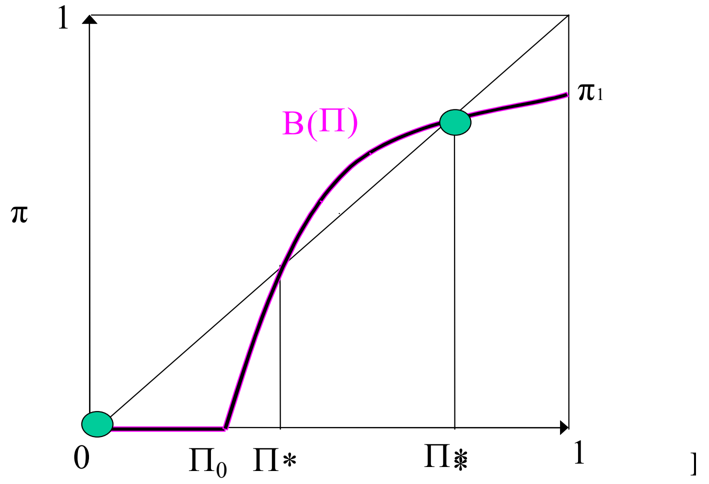


Figure 4: High Cost Case

with social trust levels in $(0, 1)$*, the stable one is unique. ii) Full trust is achievable in stable NE when condition (4) is satisfied but never so when (6) holds. iii) 'No trust' equilibrium always exists which is stable under condition (3).*

The above proposition summarizes some common results of the above four cases. It implies that low-cost players are crucial in generating positive social trust, and high-cost players in achieving full trust. There exist multiple equilibrium where up to two of them are stable. However, there is only one stable equilibrium with $\pi \in (0, 1)$. Our discussions below focus on this unique stable interior equilibrium with relevant comparative statics specified in Proposition 8.

3.4 Several Ways to Increase Social Trust

3.4.1 The Information Structure

An individual's trustworthiness sometimes can be assessed from his appearance, attitudes, and spontaneous responses (Frank 1987). It is also revealed by his actions in one-period prisoner's dilemmas (section 2). The accuracy of this encoding process often increases with one's knowledge and experiences, and decreases with the heterogeneity of partners' backgrounds. Indeed, subjects paired with a partner of a different race or nationality are less cooperative in public goods experiments (Glaeser et al. 2000), and people show lower trust in a less homogenous community (Alesina and La Ferrara 2002).

The information structure in a society is represented by p in our simple social trust formation model. Better information flow through efficient mass communication and dense social networks help facilitate the revelation of cooperative tendencies and thus lead to higher p, since more information can be accumulated about how to assess people's trustworthiness in certain circumstances and about specific individuals' behaviors. Correspondingly the amount of social trust is increased in a stable equilibrium (Proposition 8). This result helps explain the following empirical findings. Temple and Johnson (1998) show 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. 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." Putnam (1995) shows that weakened social networks may have contributed to the steady decline of social trust in US.[16]

3.4.2 Extrinsic Incentives and Intrinsic Discipline

Game-specific payoffs in prisoner's dilemmas such as d and l represent *extrinsic incentives* to defect. They are determined by disciplinary institutions including the legal system, firm incentive and monitoring schemes, social networks and social norms. The more effective these institutions are in punishing defecting behaviors, the lower d and l, which leads to higher social trust in stable equilibrium by Proposition 8.

In contrast, the cooperative tendency is an *intrinsic discipline* against defecting. The developing process of cooperative tendency is primarily conducted at home and in schools during one's childhood. When the investment costs are reduced or expected returns are increased, people would invest more in cooperative tendencies. For example, among children

[16] It is helpful to note that the information structure is quite different across countries. Social networks are dense in most developing countries so that difference in mass communication has more explanatory power. In developed countries, however, mass communication is usually effective, while the density of social networks may vary a lot.

of poor family backgrounds those who attended early intervention programs such as Headstart are more likely than others to adopt pro-social behaviors (Heckman, 1999; Garces, Duncan, and Currie, 2002). Higher time preference β and larger gain G from established cooperative matches increase benefits of being cooperative, where G is higher when cooperative partners can sustain longer tenures in families, firms, and communities (Putnam 1995).

Both extrinsic incentives and intrinsic discipline can affect people's behaviors. To achieve cooperation we can either improve the efficiency of institutions, or increase the net returns of cooperative tendency, or both. How to allocate resources between them depends on their relative costs. When the cooperative tendency is endogenous, however, their relationship is more complex than simple substitution, since outside disciplines may crowd out innate ones.[17]

Proposition 10 *Effective disciplinary institutions lead to lower average cooperative tendency.*

Proof. The average cooperative tendency in equilibrium $\underline{\alpha}= \Pi^*d + (1-\Pi^*)l$ increases with d and l because

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

Effective disciplinary institutions, by reducing d and l, induce players to choose lower $\underline{\alpha}$. ■

The intuition is that effective disciplinary institutions reduce defecting benefits and thus the threshold cooperative tendencies, which makes investment appealing to more people. As a result the proportion of cooperative people (the amount of social trust) is higher. When the disciplinary institutions are less effective, people have to invest in higher cooperative tendencies to achieve cooperation. So fewer people are cooperative, but these cooperative ones are able to withstand the temptation of larger defecting benefits. This result suggests

[17] SeeBar-Gill and Fershtman (2000), Frey, Bohnet and Huck (2001) for more examples of motivation crowding-out.

that survey-based trust indicator $TRUST$ is higher in a country with more effective disciplinary institutions, but experiment-based social trust measure may be lower if subjects are faced with quite high defecting benefits. It may account for the contradictory social trust ranking across countries. For example, $TRUST$ in the UK (44.4) is much higher than Italy (26.3) (Knack and Keefer 1997), but UK subjects "free-rode to a much greater extent" than Italians in a public goods experiment (Burlando and Hey 1997). A similar comparison is U.S. ($TRUST$=45.4) v.s. Germany ($TRUST$=29.8), where U.S. subjects free-rode more than Germans (Weimann 1994).

3.5 An Extension with Human Capital

The cooperative tendency, a trait invested in a person that yields future returns to him/her, is essentially a component of human capital.[18] It is distinct from cognitive ability h, the conventional component of human capital, in that h directly enters a specific production function, while cooperative tendency enables people to use h properly by cooperating with each other. These two components together (h, α) determine a person's overall productivity; while at investment stage, their relationship is similar to that between a child's cognitive and social development. In this section the basic social trust formation model is extended with cognitive ability investment. Since almost all previous results carry over, we only discuss differences and new findings. The proofs for the following two propositions and the human capital version of Lemma 2 are in the appendix.

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

3.5.1 Human Capital Version of the Stage Game

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

The production functions $g(\cdot), d(\cdot), l(\cdot)$ increase and are 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.

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 $\underline{\alpha}(h^i, \pi) \equiv \pi d(h^i) + (1-\pi) l(h^i)$. These two results are direct extensions of Proposition 2 and 3, respectively. Note that the threshold cooperative tendency in a game increases with a player's cognitive ability h, i.e. $\partial \underline{\alpha}(h^i, \Pi)/\partial h^i \geq 0$, since players with higher h can produce higher defecting benefits.

3.5.2 Human Capital Investment Model

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

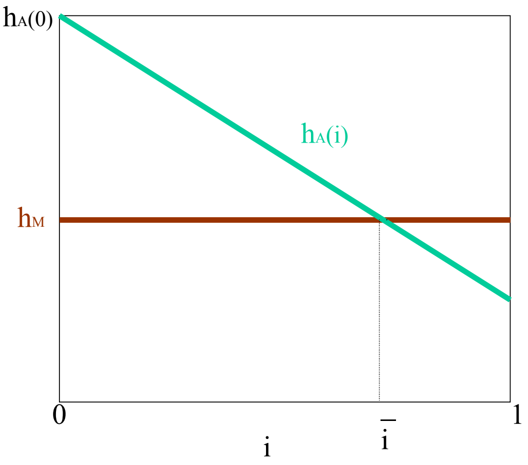


Figure 5: Relation Between h_A^i and h_M

Lemma 4 *h_A^i and h_M^i exist and are unique. h_M^i increases in Π but decreases in p. h_A^i increases with p, and also in Π if $c_{h\alpha}(h, \alpha(h, \Pi), i) \geq 0$ and $d'(h) \leq l'(h)$.*

To study the effects of cooperative tendency on cognitive ability, we assume the marginal cost of investing in h is the same across players, i.e. $c_{hi}(h, \alpha, i) = 0$. Meanwhile we maintain the same assumption $c_{\alpha i}(h, \alpha(h), i) > 0$ as before.

Proposition 11 *i) For any given Π, h_A^i strictly decreases with i, while $h_M^i = h_M$ holds for all players. ii) There exists a unique $\bar{\imath}(\Pi, p) \in [0, 1]$ such that $h_A^i \geq h_M$ for all $i \leq \bar{\imath}$, while $h_A^i < h_M$ for all $i > \bar{\imath}$. $\bar{\imath}(\Pi, p)$ increases with p and Π.*

This proposition, illustrated by figure 5, demonstrates the interaction between the two components of human capital. Players would choose the same cognitive ability h_M if not investing in cooperative tendency, since their marginal costs are the same. If they invest in cooperative tendency, lower cost ($i \leq \bar{\imath}$) players would choose higher cognitive abilities than h_M, while high cost ones ($i > \bar{\imath}$) the opposite. Therefore cognitive ability and cooperative tendency complement (substitute) each other for low (high) cost players. As p or Π goes up, they become complements for more people.

4 Conclusions

Social trust is an important social phenomenon which has been extensively studied by social scientists (see for example Cook 2001). Empirical work in the economics literature has shown that it facilitates economic performance at various levels. This paper formalizes the concepts of trustworthiness, trust, and social trust based on a single analytical element 'cooperative tendency'; and studies the formation of social trust in a society using a model of human capital investment. It provides plausible explanation for many empirical and experimental results about social trust. For example, the same player may exhibit different levels of trustworthiness across games, which in part leads to discrepancies among empirical measures of social trust (Glaeser et al. 2000); trust is lower among people with less homogenous backgrounds (Alesina and La Ferrara 2002) and it is positively correlated with mass media and social networks (Temple and Johnson 1998, Putnam 1995)); social trust level is significantly associated with economic performance (Knack and Keefer 1997), especially in large organizations where people often do not know each other well (La Porta et al. 1997); long-run social trust levels may be quite different in otherwise identical groups (Putnam 1993).

The paper also generates fresh insights and policy implications about social trust, especially on its relationship with human capital and disciplinary institutions. For instance, cultivating cooperative tendency is important to social trust and economic performance, and it may complement investment in cognitive skills. But individuals lack appropriate incentives to develop cooperative tendencies due to strong positive externalities. These results suggest that changes should be made to current human capital policies which "...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 1999), especially when the under-investment in appropriate working habits and attitudes has already affected many firms (Cappelli 1995).

Furthermore, establishing institutions to curb defecting should be optimally weighed against cultivating cooperative tendencies, taking into consideration the dynamic interactions between them. For example, criminal rates could be reduced either by more policing or by helping more children through Headstart or similar programs (Garces, Duncan, and Currie, 2002); in a firm both its incentives/monitoring scheme and the social trust among employees can increase total effort. More analysis is needed to understand the dynamic relationship between social trust and institutions.

The paper also sheds light on the relationship between various forms of social capital. For example, social networks and norms not only interact with current social trust in promoting cooperation through reputation effects, but may also affect future social trust formation. On the other hand, social trust is likely to play a crucial role in the creation and maintenance of these social capital forms as long as discrepancies between social and individual returns are involved. A more thorough treatment is left for future research.

References

1. Alesina, Alberto and Eliana La Ferrara, “Who Trusts Others?” *Journal of Public Economics*, August 2002, 85:207-34.

2. Andreoni, J. and R. Croson (2002), “Partners versus Strangers: Random Rematching in Public Goods Experiments,” forthcoming in *Handbook of Experimental Economics Results.*

3. Arrow, K. (1972), “Gift and Exchanges,” *Philosophy and Public Affairs*, I (1972), p343-62.

4. Bar-Gill, O., and C. Fershtman (2000), “The Limit of Public Policy: Endogenous Preferences,” *working paper*, Aug. 2000.

5. Becker, G. (1996) *Accounting for Tastes*, Harvard University Press.

6. Bohnet, Iris, Bruno Frey, and Steffen Huck, “More Order with Less Law: On Contract Enforcement, Trust, and Crowding,” *American Political Science Review*, Vol. 95, No. 1, March 2001: 131-144.

7. Bowles, S., and H. Gintis (1998) “The Determinants of Individual Earnings: Cognitive Skills, Personality, and Schooling,” *working paper.*

8. 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.

9. Burlando, R., and J.D. Hey (1997), “Do Anglo-Saxons Free-ride More?” *Journal of Public Economics* 64 (1997) 41-60.

10. Cappelli, P. (1995), “Is the ‘Skill Gap’ Really About Attitudes?” *California Management Review* Vol. 37, No.4, Summer 1995.

11. Coleman, J.S., “Social Capital in the Creation of Human Capital,” *American Journal of Sociology* 94 (1988): S95-S120.

12. Cook, Karen S. (2001), editor, *Trust in Society,* New York: Russell Sage Foundation, 2001.

13. Cripps, Martin, George Mailath, and Larry Samuelson (2003), "Imperfect Monitoring and Impermanent Reputations" *Econometrica*, forthcoming.

14. Dixit, Avinash, "On Modes of Economic Governance," *Econometrica*, 71(2), March 2003, 449-481.

15. Fehr, E. and S. Gachter (2000), “Fairness and Retaliation: The Economics of Reciprocity,” *Journal of Economic Perspectives* 14, 159-182.

16. Frank, Robert H. (1987), “If Homo Economicus Could Choose His Own Utility Function, Would He Want One With a Conscience?" *American Economic Review*, 77 No. 4, 593-604.

17. 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.

18. Garces, Eliana; Thomas, Duncan; Currie, Janet, “Longer-Term Effects of Head Start,” American Economic Review, September 2002, v. 92, iss. 4, pp. 999-1012.

19. Glaeser, E.L., David Laibson, and Bruce Sacerdote (2000), "The Economic Approach to Social Capital," *NBER working paper* 7728.

20. Glaeser, E.L., David Laibson, J.A. Scheinkman, and C.L. Soutter (2000), "Measuring Trust," *Quarterly Journal of Economics*, Aug. 2000.

21. Hardin, Russell. (2002), *Trust and Trustworthiness*, New York: Russell Sage Foundation.

22. Heckman, James J. (1999), "Policies to Foster Human Capital," *NBER Working Paper* 7288, August 1999.

23. James, Harvey S. (2002), "The Trust Paradox: A Survey of Economic Inquiries into the Nature of Trust and Trustworthiness," *Journal of Economic Behavior and Organization*, 47(3), pp. 291-307.

24. Kandel, E. and Lazear, E.P., 1992. "Peer pressure and partnerships," *Journal of Political Economy* 100 4, pp. 801–817.

25. 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.

26. Kreps, D.(1997), "Intrinsic Motivation and Extrinsic Incentives," *American Economic Review*, May 1997.

27. 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.

28. La Porter, R., F. Lopez-de-Silanes, A. Shleifer, and R.W. Vishny, "Trust in Large Organizations," *American Economic Review*, May 1997.

29. OECD (2001), *The Well-being of Nations: the Role of Human and Social Capital*, Paris.

30. 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.

31. Putnam, R. D. (1993) (with R. Leonardi and R.Y. Nanetti), *Making Democracy Work*, Princeton, NJ: Princeton University Press, 1993.

32. Putnam, R. D. (1995), "Bowling Alone: America's Declining Social Capital," *Journal of Democracy*, Vol.6 (1995), pp. 65-78.

33. Rob, R., and P. Zemsky (2002), "Social Capital, Corporate Culture and the Incentive Intensity," *RAND Journal of Economics*, Vol. 33 No. 2, Summer 2002.

34. Rotemberg, J. J. (1994), "Human Relations in the Workplace." *Journal of Political Economy* 102 (August 1994): 684-718.

35. Temple, J. and P.A. Johnson (1998), "Social Capability and Economic Growth," *Quarterly Journal of Economics*, August,1998.

36. Tirole, J. (1996), "A Theory of Collective Reputations," *Review of Economic Studies* (1996) 63, 1-22.

37. Weimann, J. (1994), 'Individual Behavior in a Free Riding Experiment," *Journal of Public Economics* 54, 185-200.

Appendix

- **Proof for Proposition 4.**

Proof. Given the belief system, non-selfish players would not deviate by the same arguments in Proposition 3 and 2. At period T, playing D is a selfish player's dominant strategy, so he will not deviate. If he deviates in some period $t < T$ by playing D, his selfish type is revealed. According to the equilibrium strategies, (D, D) would be played in all future periods unless his partner is selfless in which case (C, D) is played. So the deviation payoff for a selfish player from period t until T is $(g+d)\beta^{t-1}+(g+d)(\beta^t+\beta^{t+1}+...+\beta^{T-1})\pi_{SI}$. By not deviating he can get $g\beta^{t-1}+g\beta^t+...+g\beta^{T-2}+(g+d)\pi\beta^{T-1} = g\beta^{t-1}(1-\beta^{T-t+1})/(1-\beta) + d\pi\beta^{T-1}$. The non-deviation condition at period t for a selfish player is

$$[g-(g+d)\pi_S]\frac{\beta(1-\beta^{T-t-1})}{1-\beta} + (g+d)(\pi-\pi_S)\beta^{T-t} > d$$

The LHS is the net gain of cooperation at time t. Its partial derivation with respect to t is

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

which is negative if

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

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 player does not deviate at period $T-1$, then they will not deviate at any time earlier. Non-deviation at period $T-1$ means

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

It is easy to check that condition (7) is implied by condition (8) because $\pi < 1$. So that 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 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 and any period with history that only (C, C) has played, the probability of matching with a non-selfish partner is equal to π since the match is random and all behave in the same way. 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. The probability of the player who plays C in (C, D) being non-selfish is still π because both types could have done so according to the equilibrium strategy profile. ■

- **Proof for Proposition** 8.

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

$$\frac{\partial B(\Pi)}{\partial d} = \frac{\partial i^*(\Pi)}{\partial d} = -\frac{\partial V_d(i^*, \Pi)/\partial d}{\partial V_d(i^*, \Pi)/\partial i^*} = \frac{-(\beta(1-p)+c_D)\Pi}{-\partial V_d(i^*, \Pi)/\partial i^*} < 0.$$

Using exactly the same techniques, we get

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

Again by implicit function theorem, we get from $lim_{\Pi \to 0} V_d(i = \pi_0, \Pi) = \beta p G - \beta(1-p)l - c(l, \pi_0) = 0$ the following results $\frac{\partial \pi_0}{\partial d} = 0$, $\frac{\partial \pi_0}{\partial l} = -\frac{-\beta(1-p)-c_l}{-c_i} < 0$, $\frac{\partial \pi_0}{\partial \beta} = -\frac{pG-(1-p)l}{-c_i} > 0$, and $\frac{\partial \pi_0}{\partial G} = -\frac{\beta p}{-c_i} > 0$. ■

- **Proof for Proposition 9.**

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 first two cases and omitted. ■

- **Proof for Lemma 4.**

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

The objective functions are

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

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

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

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

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

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

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

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

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

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

A sufficient condition for the second order condition to hold is

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

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

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

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

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

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

if $\frac{\partial c(h,\alpha(h,\Pi),i)}{\partial h \partial \Pi} \leq 0$ holds. Given that

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

a sufficient condition is

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

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

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

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

- **Proof for Proposition 11.**

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 use the Implicit Function Theorem and get

$$\frac{\partial h_M^i}{\partial i} = -\frac{\partial [V_M^i(h)]'}{\partial i} / \frac{\partial [V_M^i(h)]'}{\partial h} = \frac{\partial c_h(h, 0, i)}{\partial i} / \frac{-\partial^2 V_M^i(h)}{\partial h^2} \gtreqless 0,$$

iff $\partial c_{hi}(h, 0, i) \lesseqgtr 0$. Similarly from $[V_A^i(h)]' = 0$ we get

$$\frac{\partial h_A^i}{\partial i} = -\frac{\partial [V_A^i(h)]'}{\partial i} / \frac{\partial [V_A^i(h)]'}{\partial h} = \frac{\partial^2 c(h, \alpha(h,\Pi), i)}{\partial h \partial i} / \frac{\partial^2 V_A^i(h)}{\partial h^2} \gtreqless 0,$$

iff $\partial^2 c(h, \alpha(h,\Pi), i)/\partial h \partial i = c_{hi}(h, \alpha(h,\Pi), i) + c_{\alpha i}(h, \alpha(h,\Pi), i)\alpha_h(h,\Pi) \lesseqgtr 0$. Under the following assumption,

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

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

(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 $[V_A^i(h_M)]' \gtreqless 0$, then $h_A^i \gtreqless h_M$ is proved. By condition (9), $[V_M^i(h)]' = 0$, which is $-\beta\Pi(1-p)[g'(h_M) + d'(h_M)] + c_h(h_M, 0, i) = 0$. Add this zero term to $[V_A^i(h_M)]'$, we get

$$\begin{aligned} [V_A^i(h_M)]' &= \underbrace{\beta p G'(h_M) - \beta(1-p)[\Pi d'(h_M) + (1-\Pi)l'(h_M)]}_{A(\Pi)} \\ &\quad -\underbrace{[\partial c(h_M, \alpha(h_M, \Pi), i)/\partial h_M - c_h(h_M, 0, i)]}_{B(i,\Pi)}. \\ &\equiv A(\Pi) - B(i, \Pi). \end{aligned} \tag{11}$$

The first term $A(\Pi)$, the gain of investing in α, is the same for all players. The second term $B(i,\Pi)$ is the investing cost for player i, which increases with player index since

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

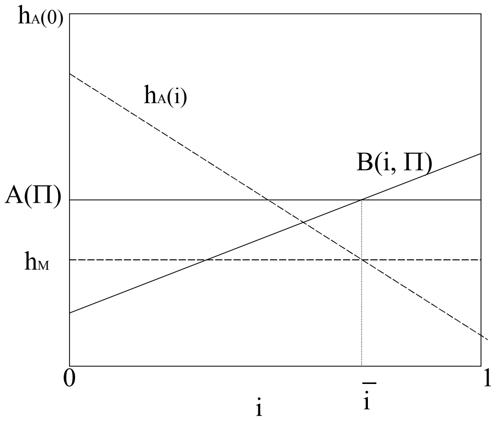
If the boundary condition $[V_A^0(h_M)]' \geq 0 \geq [V_A^1(h_M)]'$ holds given Π and p, i.e. if

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

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

$$[V_A^{\bar{\imath}}(h_M)]' = A(\Pi) - B(\bar{\imath}, \Pi) = 0. \tag{12}$$

This condition means that for player $\bar{\imath}$, his optimal choice $h_A^{\bar{\imath}}$ is equal to h_M, not affected by his choice of α. Then for all $i \leq \bar{\imath}$, we have $[V_A^i(h_M)]' > 0 \iff h_A^i \geq h_M$; for all $i > \bar{\imath}$, $[V_A^i(h_M)]' < 0 \iff h_A^i < h_M$. 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. See figure below for illustration.



Now we check the sign of $\frac{\partial \bar{i}}{\partial \Pi}$ based on equation (12).

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

Similarly we have $\partial \bar{i}(\Pi, p)/\partial p = \beta^2 G'(h_M)/c_{\alpha i}(h_M, \alpha(h_M, \Pi), i)\alpha_h(h_M, \Pi) > 0$. ■

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

Proof. By the Envelope Theorem,

$$\begin{aligned} \frac{\partial V_d(i, \Pi)}{\partial i} &\equiv \frac{\partial V_A^i(h_A^i) - \partial V_M^i(h_M)}{\partial i} = -[c_i(h_A^i, \alpha(h_A^i, \Pi), i) - c_i(h_M, 0, i)] < 0. \\ \frac{\partial V_d(i, \Pi)}{\partial \Pi} &= \beta(1-p)[g(h_A^i) \ + l(h_A^i) \ - g(h_M) - d(h_M)] \\ &\quad + c_\alpha(h_A^i, \alpha(h_A^i, \Pi), i)[l(h_A^i) - d(h_A^i)] \end{aligned}$$

It is obvious that $\partial V_d(i, \Pi)/\partial \Pi > 0$ when $h_A^i \geq h_M$, which is true for low index players. If we can show that $\frac{\partial V_d(i,\Pi)}{\partial \Pi}$ decreases with player index, then $\partial V_d(i, \Pi)/\partial \Pi > 0$ for all players. Indeed this is the case since $\frac{\partial^2 V_d(i,\Pi)}{\partial \Pi \partial i} = c_{i\alpha}(h_A^i, \alpha(h_A^i, \Pi), i)[l(h_A^i) - d(h_A^i) + \alpha_h \partial h_A^i/\partial \Pi] > 0$. The intuition is that high cost players get more benefits from reduced $\alpha(h_A^i, \Pi)$ due to a higher Π. ■