On the Reliability of Trusting

Harvey S. James, Jr.[*]

Agribusiness Research Institute
University of Missouri-Columbia
124 Mumford Hall
Columbia, MO 65211-6200
Phone: 573-884-9682
Fax: 573-882-3958
E-mail: hjames@misouri.edu

November 20, 2001

Abstract: This paper presents a model of trust in which a principal chooses either to trust or monitor an agent who, in turn, chooses either to honor or exploit that trust. The principal's decision of whether to trust or monitor is based on the *relative temptation* an agent faces to exploit the principal's trust, which comprises two elements – the environmental incentives the agent faces and the personal characteristics of the agent. The model is used to develop a reliability condition that the principal uses to assess the likelihood that trust placed in an agent will be honored.

Keywords: trust, trustworthiness, principal-agent relationship, moral hazard, transaction costs, monitoring

---

[*] I am grateful for comments from Kenneth Koford, Farhad Rassekh, and anonymous referees, which greatly improved the exposition of this paper. An earlier version was presented at the annual meetings of the Eastern Economic Association.

On the Reliability of Trusting

Introduction

Trust is an important aspect of economic exchanges, because it is a part of every transaction and because it reduces transaction costs and thus has efficiency consequences. Arrow, for instance, states that "there is an element of trust in every transaction" (1973, p. 23) and that trust is both necessary and sufficient for economic activity (1974). When parties to an exchange trust each other – that is, when each expects the other to perform an agreed-upon action – they may not necessarily have to rely on costly contracting, monitoring, and enforcement schemes to guide their behavior. As important trust is to economic activity, however, the theoretical literature on trust formation is only in its infancy (see Dasgupta, 1988).

Kreps (1990) introduces a one-sided variation of the prisoner's dilemma he calls the "trust game" (see Figure 1). In this game an agent, A, chooses either to trust or not trust another agent, B, who, if trusted, chooses either to honor that trust or abuse it. In a one-shot play of the game B has an incentive to abuse A's trust, inducing A not to trust. The insight of Kreps' model is that A will trust if B has a reputation for trustworthiness, which can develop if the interaction between A and B is repeated with sufficient probability. For this reason, some scholars have approached the problem of trust formation from an evolutionary perspective (see Güth and Kliemt, 1994). Lahno (1995) builds on this theme by developing a model of trust based on a linear recursive function of reputation formation. With this model Lahno shows that an agent will trust another agent not only when the reputation of the second is sufficiently strong but also, under some circumstances, when that reputation is low. That is, trusting and trustworthiness are not subgame perfect strategies because moves off the equilibrium path by a trustee who exploits trust could be restored under certain conditions of reputation formation. Though insightful,

these models of trust presuppose a history of play upon which reputations are built. Absent time, trust is difficult to explain.

The model developed by Coleman (1990) overcomes this problem. In his model agent A trusts agent B when the expected gains from trusting exceed the expected losses. Specifically, if π is the probability that B is trustworthy, G is the potential gain to A if B is trustworthy, and L is the potential loss to A if B is untrustworthy, then A will trust when $\pi G > (1-\pi)L$, or when $\frac{\pi}{1-\pi} > \frac{L}{G}$. In this model, a reputation built on repetition is neither necessary nor sufficient for trusting. For example, two agents may have strong reputations for trustworthiness, though one agent may be trusted and the other not depending on the ratio of losses to gains resulting from trust placed in each agent. The weakness of Coleman's model is that it neither makes explicit the losses and gains, nor does it fully explore the non-cognitive or psychological factors of the trustor and trustee that affect the formation of trust. For instance, if the agent in whom trust is placed feels remorse for exploiting the trust of another, and if the trustor understands this, then trust could be fostered under certain cultural (Huang and Wu, 1994) or legal (Huck, 1998) environments.

Snijders and Keren (1999) present a model in which the trust game of Kreps (1990) is used to make explicit the loss-gain ratio of Coleman (1990) by incorporating specific psychological states of both the trustor and trustee. Suppose an agent, A, has utility $\mathrm{U}(m)$, where m is the monetary payoff resulting from the play of Kreps' trust game from Figure 1, and π is the probability that B will honor A's trust. Snijders and Keren show that A will trust B when $\pi\,\mathrm{U}(x) + (1-\pi)\,\mathrm{U}(z) > \mathrm{U}(y)$, or when $\pi > \frac{\mathrm{U}(y)-\mathrm{U}(z)}{\mathrm{U}(x)-\mathrm{U}(z)}$. The numerator represents the potential loss from misplaced trust, while the denominator represents the potential gain from trusting. Snijders and Keren use this relationship to show how trust is affected by non-cognitive components, such as social conditions, guilt felt by the trustee for exploiting trust, and regret felt by the trustor for misplacing trust.

These models provide an important foundation for the study of trust. However, they suffer from several weaknesses. First, the models do not incorporate an alternative to trust – when an agent chooses not to trust the game ends. Clearly, alternatives to trusting exist, such as insurance, monitoring, sanctions, rewards, litigation, or combinations of these (Leibenstein, 1987), although they may be imperfect substitutes. Second, though Coleman (1990) and Snijders and Keren (1999) interpret the probability π as A's subjective evaluation that B will be trustworthy, it is exogenous to the model. The problem with this approach is that there is no explanation as to how π is determined and how it is affected by information or other factors A uses to assess the likelihood that trust placed in B will be honored. Third, these models depend on the premise that A trusts B only when A expects B to be sufficiently trustworthy, resulting in the obvious explanation of why A trusts B (see James, 2002) – if the probability that B is trustworthy exceeds some minimal threshold then A will trust B, otherwise A will not trust B. A more interesting question is whether there are instances in which A would trust a relatively untrustworthy B or not trust a B who is sufficiently trustworthy. Accordingly, one must wonder whether there is a more reliable method of modeling trust that incorporates these possibilities.

The purpose of this paper is to extend Coleman's (1990) framework by modeling trust in a principal-agent relationship in which the alternative to trusting is a costly action taken by a principal to observe and verify an agent's performance. In this model a decision not to trust is a decision to monitor rather than quit, where monitoring is recognized as one imperfect, though plausible, substitute for trust. To be precise, a principal "trusts" an agent if she assigns a task to him and then allows him to act without direct monitoring or supervision. If the principal "mistrusts" the agent, she takes a costly action to observe and verify his performance. The model also integrates the dual possibilities that a principal may fail to trust a (genuinely) trustworthy agent or mistakenly trust an untrustworthy one by examining a reliability condition that the principal must satisfy before trusting an agent, even if the agent is known to have a sufficiently high probability of trustworthiness. The reliability condition is based on the premise that the decision of whether to trust or monitor is based on a concept called *relative temptation*, and it reflects the

extent to which an agent may exploit the principal's trust. An implication is that an agent may in fact be trustworthy in any and all circumstances but still not be trusted by a principal because of an impression or perception held by the principal.[1] Conversely, a principal may recognize that extensive incentives exist for an agent to shirk, but still choose to trust the agent because the principal perceives the agent to have sufficiently high qualities of character to be trustworthy in that context.

The relative temptation of the agent to exploit a principal's trust is assumed to be affected by two major classes of variables. The first consists of environmental factors that affect an agent's rational self-interest only, including formal rules, informal social norms, and enforcement mechanisms (see North, 1990). An example would be a one-shot play of the Prisoner's Dilemma in which the rational strategy of the players is to defect rather than cooperate. This idea reflects the amoral *Homo economicus* – the calculating, rational, self-interested decision-maker who chooses actions based solely on salient incentives without regard for moral implication. Thus, *Homo economicus* is honest or trustworthy *only* if he has an incentive to be.[2] The second class of variables defines the dispositions, proclivities and other qualities of the agent's character.[3] This idea is based on *Homo ethicus* – a fully altruistic individual (or willing "martyr"), who sacrifices personal pleasure for the good of others or fulfills his duty even at great personal expense. Thus, *Homo ethicus* will be honest in spite of the incentives he faces. Hence, in a one-shot or finitely repeated play of the Prisoner's Dilemma, players who cooperate would have higher qualities of characters than those who do not cooperate.[4] The relative temptation of the agent to exploit the trust of another is represented as a gap between the incentives the agent faces and the agent's personal

---

[1] Thus, first impressions are important, even in matters of trust. (See Quigley-Fernandez, Malkis, and Tedeschi, 1985.)

[2] Consistent with economic expectations, Snijders and Keren (1999) present experimental evidence showing that the greater is an agent's gain from exploiting the trust of another (i.e., the greater the economic temptation to exploit trust), the less likely that agent will honor trust. See Hardin (1993) and James (2002) for related discussions.

[3] Regarding the various aspects of these qualities, Kay (1996) notes that the "qualities required of the [agent] would appear to be virtuous (loyalty, veracity, etc.) and would appear to imply that associated actions would be offered at some cost to the individual or individuals without regard to any obvious compensatory gains on their part (otherwise the actions would tend to reduce to rational self-interest)" (p. 252).

[4] Since research suggests that economists are generally less willing to cooperate than non-economists (Frank, Gilovich, and Regan, 1993), one may argue that economists have lower qualities than others. See Cohen and James

characteristics. In general, the stronger are the environmental incentives for an agent to exploit trust, or the weaker are the agents' moral qualities, other things being equal, the greater will be the relative temptation the agent faces.

The model of trust developed in this paper builds on the general problem of moral hazard in principal-agent relationships, which arises when observing and verifying agent performance is costly. In these cases agents have an incentive to exploit the information asymmetry for personal advantage (Jensen and Meckling, 1976; Holmstrom, 1982). Accordingly, principals seek to design contracts that best align the interests of the agent with those of the principal. Typically, principal-agent models examine various combinations of incentives, monitoring, and job design as solutions (see Holmstrom and Milgrom, 1991). This model differs from other principal-agent models in that it examines trust rather than compensation or job design as an alternative to monitoring. In this model trust is affected by the principal's perception of the relative temptation the agent faces to act in his, rather than the principal's, interest. Moreover, the idea that both environmental and personal quality characteristics affect the willingness of a principal to trust an agent is in contrast with the transaction cost literature that says environmental factors only, such as the cost of monitoring, principally determine the likelihood that a principal relies on observation, verification or other contracting or governance schema in the structure of exchange relationships (see James, 2000), in part because agents are assumed always to behave opportunistically.[5] According to the transaction cost framework, when transactions are subject to *ex post* opportunism they would benefit from *ex ante* safeguards, such as improved governance structures or realigned contractual incentives. As this paper shows, this is only partially true. The potential for opportunism could be high, but it may also be tempered by other factors, such as known qualities of the person in whom trust is placed, thus making contracting safeguards ubiquitous even in the case of potential opportunism. This paper suggests that verification may not necessarily occur when monitoring and other transaction costs are low and that

---

(2001) for experimental evidence that ethics training can increase the likelihood that economic students will cooperate, suggesting that their character qualities increase as a result of the exposure to ethics training.

[5] This is why Williamson states that "contracting would be ubiquitous in the face of nonopportunism" (1985, p. 66).

verification may occur when monitoring costs are high, the distinguishing factor being the relative degree of temptation confronting the agents.

The Principal-Agent Model

The following model describes a principal-agent relationship in which a principal chooses either to trust that an agent will complete a contracted action or to engage in costly observation and verification of the agent's performance, and the agent chooses either to honor the contract or to shirk. The model is used to develop a reliability condition of trusting based on Heiner's (1983) model of uncertainty and the predictability of individual behavior. The reliability condition is based on the principal's perception of the relative temptation the agent faces to shirk if trusted. The idea is that as long as the reliability condition meets or exceeds a certain level, the principal will be willing to trust the agent, even if the probability of agent trustworthiness is low, in cases of potential opportunism of the agent or when the potential losses from misplaced trust are large. However, if the reliability condition falls below a certain limit, then the principal will instead rely on costly verification procedures to ensure that the agent fulfills the agreed-upon contractual promise.

The Principal-Agent Relationship

A risk-neutral Principal hires an Agent from a pool of potential agents to supply some action, $a^*>0$. The Principal derives revenue, R, per unit of action taken by the Agent, and pays the Agent a fixed wage, $w>0$, for the contracted action a^*.[6] The utility of the Agent is based on Koford's (2000) model of agency and is defined as

[6] For simplicity, it is assumed that the principal has already solved the problem of how to set the wage because the model does not consider the tradeoff between fixed and incentive wages. Nevertheless, and without loss of generality, it could also be assumed that the wage selected is the minimum wage necessary to induce the Agent to accept the contract.

$$U_A = \mathrm{U}(w) - \mathrm{V}(a) - H(a^* - a),$$

where $\mathrm{U}(w)$ is the utility of the wage, $\mathrm{V}(a)$ is the disutility of effort taken, and *H(a*-a)* is "the disutility associated with violating the contractual promise, or 'shirking,' which occurs when the action taken, *a*, is not equal to the contracted action, *a**, where the * indicates the contractual value" (p. 3). The functions $\mathrm{U}(\cdot)$ and $\mathrm{V}(\cdot)$ in the Agent's utility are assumed to be the same across all potential agents so that they have identical preferences for wealth and identical abilities. For simplicity, the utility of the wage is normalized so that $\mathrm{U}(w) = w$, and $\mathrm{V}(a)$ is assumed to be strictly increasing and convex. Note that because the wage paid by the Principal to the Agent is fixed, the Agent has no incentive to supply $a > a^*$. Though potential agents are identical in abilities and preferences for wealth, they vary in terms of their honesty and diligence in working as contracted, so that higher values for the number *H* represent more honest workers and $H \geq 0$.[7] By definition, the Agent "behaves honestly" if he supplies effort $a = a^*$, while the Agent "shirks" if he supplies effort $a < a^*$. Thus, when the Agent is contracted to supply effort for the Principal, the Agent has two options: complete the agreed-upon action or shirk.

Assume further that the Principal neither knows perfectly the *H*-type of the Agent nor observes costlessly what action he takes. The Principal can detect shirking with probability $0 < p < 1$ by taking an action at cost $c(p) > 0$ to observe and verify the Agent's performance. If the Agent is caught shirking, he is fired by the Principal without pay, but doing so also produces no revenue for the Principal. The cost of observing and verifying is strictly increasing and convex. Because the Principal cannot perfectly detect the *H*-type and action of the Agent, the Principal has two options. First, she could trust that the Agent will perform the contracted action, in which case the Principal chooses only the contracted level of effort, *a**. Second, she could monitor at cost $c(p)$ rather than trust, in which case she chooses not only *a** but also the probability *p* of detecting shirking (and thus firing a shirking worker).

---

[7] The terms "honesty" and "trustworthiness" in an agent are often used interchangeably, but they are not equivalent because trustworthiness can be understood as a subset of honesty. In this paper, a trusted agent who behaves honestly by completing the agreed-upon action is said to be trustworthy, while a monitored agent who completes the action, by definition, is not trustworthy because trust is not placed in him.

The Principal-Agent relationship can be modeled in part as an extensive form game, which is illustrated in part in Figure 2. The timing of the model is as follows: First, the Principal randomly selects an Agent from a pool of agents with known identical abilities who vary only in terms of their *H*-types,[8] and offers the Agent a contract in which he is to perform action *a** in exchange for wage, *w*, which the Agent accepts so long as $U_A \geq 0$. Second, the game follows Figure 2 in that the Principal chooses either to trust the Agent or monitor with policy *p*, which probability will be known by the Agent if the Principal monitors. Third, the Agent chooses either to work as contracted or shirk. Fourth, uncertainty ends and payoffs to the Principal and Agent are realized.[9]

Trusting by the Principal. Suppose the Principal trusts the Agent (denoted by *T* in the subscript of the Agent's utility). If the Agent behaves honestly (superscript *h*) and supplies the contracted effort, then his utility is $U_T^h = w - \mathrm{V}(a^*)$. However, if the Agent shirks (superscript *s*), then the Agent's utility is $U_T^s = w - \mathrm{V}(a) - H(a^* - a)$. The agent will honor the contract when $U_T^h \geq U_T^s$. The value of H that is the boundary between honest and shirking performance is denoted by $\tilde{H}$. Equating U_T^h and U_T^s and solving results in

$$\tilde{H} = \frac{\mathrm{V}(a^*) - \mathrm{V}(a)}{a^* - a}.$$

If the Agent's *H*-type is greater than or equal to $\tilde{H}$, his utility is maximized by performing the contracted action, *a**. If the Agent's *H*-type is less than $\tilde{H}$, his utility is maximized by shirking such that $a < a^*$. Because the Agent is selected from a pool of agents with randomly distributed *H*-types, define

[8] To be clear, *a priori*, the Principal knows the abilities and preferences for wealth for the agents and the distribution of the agents' *H*-types, but not the specific *H*-type of the Agent selected, because the agents in the pool are identical except for their *H*-types.

[9] To say that uncertainty ends means that the action, *a*, taken by the Agent is revealed, either because the worker completes action *a* (when the Principal trusts or the Agent does not shirk if monitored), or because the Principal detects shirking and fires the worker, with probability *p*, if the Principal monitors and the Agent shirks.

the *a priori* probability that the Agent will complete the action *a** when trusted as $\pi(H \geq \tilde{H})$.[10] Note that $\frac{\partial \pi}{\partial a^*} < 0$;[11] the probability that the Agent will perform the contracted action decreases as the contracted action becomes more difficult. Additionally, as the minimum *H*-type for honesty increases, the Agent will be less likely to perform as contracted,[12] but if the specific *H*-type of the Agent increases, then the probability increases.

The profit to the Principal for trusting is the benefit received from the Agent's effort, less the wage paid. Specifically, if the Principal trusts and the Agent works as contracted, then the profit to the Principal is $\Pi_T^h = Ra^* - w$. However, if the Principal trusts and the Agent shirks, then the Principal's profit is $\Pi_T^s = Ra - w$.

Monitoring by the Principal. Suppose the Principal does not trust the Agent but rather monitors performance by selecting a costly detection policy of $p>0$ (denoted by *M* in the subscript of the Agent's utility). If the Agent behaves honestly (superscript *h*) and supplies the contracted effort, then his utility is $U_M^h = w - \mathrm{V}(a^*)$. However, if the Agent shirks (superscript *s*), then his utility is $U_M^s = (1-p)w - \mathrm{V}(a) - H(a^* - a)$. The Agent will honor the contract when $U_M^h \geq U_M^s$. Solving for $\hat{H}$ to identify the boundary between honest and shirking workers when the Principal monitors performance results in

$$\hat{H} = \frac{\mathrm{V}(a^*) - \mathrm{V}(a) - wp}{a^* - a}.$$

---

[10] The Principal therefore knows this probability *a priori*, or at the time the Principal randomly draws an Agent from the pool of agents. Note also that the probability that the Agent will shirk when trusted is $(1-\pi)$.

[11] See Appendix I for proof.

[12] This is true by definition: $\pi(H \geq \tilde{H}) > \pi'(H \geq \tilde{H}')$ for all $\tilde{H} < \tilde{H}'$.

Define the probability that the Agent will complete the contracted action, when monitored, as $\phi(H \geq \hat{H})$. Observe that $\frac{\partial \phi}{\partial a*} < 0$, $\frac{\partial \phi}{\partial w} > 0$, and $\frac{\partial \phi}{\partial p} > 0$.[13] The probability that a monitored Agent behaves honestly decreases as the contracted action becomes more difficult; but it increases as the wage offer increases, and the probability of detection increases, other things being equal. Furthermore, $\pi < \phi$ for any $w > 0$ and $p > 0$.[14] In other words, a monitored agent is more likely to behave honestly than one who is trusted.

If the Principal monitors the Agent's performance and the Agent works as contracted, then the Principal's profit is $\Pi_M^h = Ra* - c(p) - w$. However, if the Agent shirks, the Principal's profit is $\Pi_M^s = -pc(p) + (1-p)(Ra - c(p) - w)$. If the Principal detects shirking, with probability *p*, then the Principal terminates the agency contract without paying the Agent but also without receiving revenue, *R*. This results in an expected cost of $pc(p)$. However, with probability $(1-p)$ the Principal does not detect shirking, producing an expected return of $(1-p)(Ra - c(p) - w)$.

Standard Equilibrium Analysis

In a standard equilibrium analysis, the Agent will honor the contract when his *H*-type exceeds $\tilde{H}$, expected (*a priori*) by the Principal with probability π, and the Principal will be inclined to trust the Agent when the expected return from trusting exceed the expected return from monitoring. The expected return to the Principal from trusting is $\Pi_T^h \pi + \Pi_T^s (1-\pi)$, and the expected return from monitoring is $\Pi_M^h \phi + \Pi_M^s (1-\phi)$. Therefore, the Principal will trust when

$$\Pi_T^h \pi + \Pi_T^s (1-\pi) \geq \Pi_M^h \phi + \Pi_M^s (1-\phi).$$

---

[13] See Appendix I for proofs.

[14] $\pi(H \geq \tilde{H}) < \phi(H \geq \hat{H})$ because $\tilde{H} > \hat{H}$.

To find the minimum probability of Agent trustworthiness that is necessary to induce trust by the Principal, solve for π, perform the necessary substitutions, and simplify to obtain

$$\pi \geq \frac{R(a^* - a)\phi - (Ra - w)(1-\phi)p - c(p)}{R(a^* - a)} \equiv \pi^* .$$

This condition says that the Principal will be inclined to trust rather than monitor the Agent when $\pi \geq \pi^*$, and it is similar to the trusting conditions offered by Coleman, Snijders and Keren, and others who show that trust arises when the probability of trustworthiness exceeds some threshold. Although this approach provides some insight into the analysis of trust, it also makes trusting by the Principal trivially obvious.[15] The Principal trusts solely because the probability of trustworthiness of the Agent exceeds the threshold and not because of a conscious decision to trust or monitor. Indeed, without further refinement the model fails to allow for the possibility that the Principal may actually monitor the Agent even though the trustworthiness condition is satisfied (e.g., $\pi \geq \pi^*$) or that the Principal may trust the Agent even though $\pi < \pi^*$, the distinguishing condition being how reliable the Principal believes her *assessment* is of the Agent in this particular context. Unlike the models of Coleman and others that interpret π as a subjective probability but do not explain how the Principal incorporates information about the Agent, the analysis that follows takes π as an *a priori* probability that the Agent is trustworthy and then contrasts that with a subjective evaluation of the likelihood that the specific Agent selected will honor or violate the trust offered.[16]

Incentives, Character and Relative Temptation

Let **e** represent the class of environmental variables that determine the incentives the Agent faces to shirk or work honestly when trusted. For instance, **e** reflects the difficulty of the contracted action, the

[15] Williamson (1993) claims that trust in this context as "transparently calculative" and is thus not trust at all.
[16] An additional advantage of distinguishing between an *a priori* probability of trustworthiness and a subjective assessment is that it allows for the examination of a tradeoff between the costs of *ex ante* and *ex post* investments in

wage offer and the intensity of monitoring. An increase in **e** means that the incentives for the Agent to exploit the Principal's trust have increased, perhaps because the difficulty of the contracted action is larger than expected, for example. Let **q** represent the class of variables that describe the qualities of the Agent's character, such as the Agent's *H*-type or level of trustworthiness. An increase in **q** means that the disposition of the Agent to behave honestly is improved; that is, an Agent with a higher *H*-type will be less willing to shirk than one with a lower *H*-type, other things being equal.[17] Together, the variables **e** and **q** determine the *relative temptation* the Agent faces to exploit the trust of the Principal. Conceptually, the relative temptation the Agent faces to exploit trust may be represented as a gap between **e** and **q**. The relative temptation faced by the Agent increases when the gap between **e** and **q** increases – that is, when **e** increases for a given **q**, or when **q** decreases for a given **e**. The idea here is that the Agent may be tempted to exploit the trust of the Principal when **e** is high. Whether or not the Agent succumbs to the temptation depends on **q**. If **q** is sufficiently high, then the Agent may not exploit the trust of the Principal, even though it could be in his immediate interest to do so (for instance, if exploitation or dishonesty could occur without detection and punishment).[18] Similarly, if **q** is low but the incentives **e** are also low, then the relative temptation of the Agent to exploit trust would also be low.

The relationship between the environmental incentives and personal quality characteristic can be represented by the function, $\mathbf{X} = x(\mathbf{e},\mathbf{q})$, which defines the Principal's *perception* of the Agent's relative temptation to shirk. For example, one possible though simple expression for **X** is $\mathbf{X} = e^{\tilde{H}-H}$, where $\tilde{H}$ embodies the elements of **e** and *H* is an element of **q**. This expressions says that **X** increases as the

---

information acquisition on Agent characteristics. Although this model does not explore this tradeoff directly, it is explored briefly later in the paper.

[17] Levi (2000) identifies the factors of motives, interests, and competence as affecting trust or distrust, which would be reflected in the variable **q**. For example, whether the Agent has friendly, indifferent, or hostile motives is a characteristic of the Agent, where a friendly motive is manifested as an increase in **q** while a hostile motive is manifested as a decrease in **q**. Similarly, whether the Agent has similar or individually distinct interests from the Principal would be reflected in **q**, where similar interests correspond to a higher **q** than distinct interests. Finally, the more competent the Agent feels in being able to complete a given contracted action, the higher the level of **q** will be.

[18] Dasgupta (1988) recognizes this when he observes that "being able to trust a person to do what he said he would … requires us to know not only something of his disposition, but also something of the circumstances surrounding

Principal's assessment of the Agent's *H*-type relative to $\tilde{H}$ declines. Although the Principal knows the general probability π that the randomly selected Agent would behave honestly if trusted, the Principal's assessment of the *H*-type of the specific Agent chosen for a given degree of difficulty of the contracted activity is embodied by **X**. In essence, **X** represents the degree of uncertainty the Principal feels that the Agent selected will perform the contracted action if trusted.[19] With this formulation, **X** increases when the Principal is less certain that the Agent will complete the contract as agreed, but it decreases when the Principal is more confident that the Agent will honor the contract if trusted. Stated in general terms, **X** increases as the **e**-**q** gap increases (e.g., as **e** increases or **q** decreases, other things equal), and **X** decreases as the **e**-**q** gap decreases. Hence, the weaker are the environmental incentives (**e**) for the Agent to be untrustworthy, or the higher is the quality of the Agent's character (**q**), the smaller is the relative temptation of the Agent to shirk as perceived by the Principal and thus the greater will be the likelihood that the Principal would trust the Agent.[20]

The Reliably Trusting Condition

If π is the probability that a trusted Agent will honor the contract, then let $r(\mathbf{X})$ represent the conditional probability that the Principal correctly trusts the Agent when he is willing to honor that trust. To be clear, observe that the probability that the Principal correctly trusts is not the same as the probability that the Agent is trustworthy. Although trust and trustworthiness are often conflated (see Glaeser et al, 2000), trust refers to a choice of the Principal and being trustworthy refers to a choice of the Agent when trusted. Note that $r(\mathbf{X})$ declines as **X** increases; that is, as the Principal's assessment of the

---

the occasion at hand. If the incentives are 'right,' even a trustworthy person can be relied upon to be untrustworthy" (p. 54).

[19] Although how **X** changes is not explicitly defined in this model, one possibility is that **X** changes according to Bayesian learning, in the sense that experiences over time change the Principal's assessment of **X** (see Hardin, 1993, for a related discussion).

relative temptation facing the Agent increases, the likelihood that the Principal correctly trusts the Agent will decline. The associated gain, *G*, to the Principal is the profit, Π_T^h, from trusting an honest Agent, or

$$G = Ra^* - w .$$

Similarly, if $(1-\pi)$ is the probability that the Agent, if trusted, will shirk, then let $w(\mathbf{X})$ be the conditional probability that the Principal incorrectly trusts a shirking Agent, where $w(\mathbf{X})$ increases as $\mathbf{X}$ increases. In this case the loss, *L*, to the Principal from misplaced trust is

$$L = R(a^* - a) + \{(Ra^* - c(p) - w)\phi + [-pc + (1-p)(Ra - c(p) - w)](1-\phi)\} .$$

The loss associated with trusting a shirking worker consists of two parts. The first is the residual loss the Principal experiences because the trusted Agent shirks, equal to $\Pi_T^h - \Pi_T^s$. The second is the opportunity cost to the Principal of trusting rather than adopting a policy of observing and verifying the Agent's performance. The opportunity cost of trusting is the expected profit from monitoring, equal to $\Pi_M^h \phi + \Pi_M^s (1-\phi)$. When simplified, the expression for *L* is

$$L = R(a^* - a)\phi - (Ra - w)(1-\phi)p + Ra^* - c(p) - w .$$

When can the Principal reliably trust? Following Heiner (1983), the answer is when the Principal is confident enough that trusting will result in relatively more gains than losses. Specifically, if $r(\mathbf{X})$ is the likelihood that a Principal correctly trusts when the Agent is trustworthy, π is the probability that the Agent honors the contract, and *G* is the Principal's gain from correctly trusting, then the expected gain from trusting the Agent is $r(\mathbf{X})G\pi$. However, if $w(\mathbf{X})$ is the likelihood that the Principal mistakenly trusts the Agent who shirks, $(1-\pi)$ is the probability that the Agent shirks when trusted, and *L* is the resulting loss, then the expected loss from mistrusting an Agent is $w(\mathbf{X})L(1-\pi)$. Accordingly, the Principal will want to trust when the expected gains at least equal, or exceed, the expected losses, or when $r(\mathbf{X})G\pi > w(\mathbf{X})L(1-\pi)$. Rearranging, we obtain the following *Reliably Trusting Condition* (RTC):

---

[20] Along these lines, Baier (1986, p. 254) notes that "trusting can continue to be rational, even when there are such unwelcome suspicions, as long as the trustor is confident that in the conflict of motives within the trusted the

$$\frac{r(\mathbf{X})}{w(\mathbf{X})} \geq \frac{L(1-\pi)}{G\pi},$$

or

$$\frac{r(\mathbf{X})}{w(\mathbf{X})} \geq \frac{[R(a^*-a)\phi-(Ra-w)(1-\phi)p+Ra^*-c(p)-w](1-\pi)}{(Ra^*-w)\pi}.$$

The left hand side of the inequality represents the RTC, and the right hand side represents the minimal *Trusting Limit* (TL) that the RTC must satisfy. The RTC, which is the ratio $\frac{r(\mathbf{X})}{w(\mathbf{X})}$, represents the chance of the Principal correctly trusting when the Agent is trustworthy relative to the chance of mistakenly trusting when the Agent will shirk.[21] Note that the RTC is based on the Principal's assessment **X** of the **e**-**q** gap or relative temptation of the Agent to shirk when trusted. An increase in **X** will both reduce the chance of correctly trusting and increase the chance of mistakenly trusting, thus causing the RTC to drop. In other words, the greater is the Agent's relative temptation to exploit trust, as perceived by the Principal, the less reliable the Principal will feel in trusting, other things being equal, which will be manifested by as a decline in the RTC. The TL, on the other hand, is expressed in terms of the expected losses from misplaced trust relative to the expected gains of correctly trusting, and it determines how likely trust placed in an Agent would have to be before the Principal would want to trust.[22] The reliability condition states that the Principal should trust when the reliability of trusting (the RTC) meets or exceeds the minimum trusting limit (the TL). On the other hand, the Principal should not trust, but rather should adopt a costly policy of monitoring the Agent's performance, if the reliability of trusting is less than the minimum ratio of expected losses to gains from trusting.

---

subversive motives will lose to the conformist motives."

[21] Heiner (1983), in a footnote, observes that the "probabilities *r* and *w* can also be interpreted using Type 1 and Type 2 errors in statistical hypothesis testing." In a Type 1 error, the Principal fails to trust the Agent when in fact the Agent is willing to supply the agreed-upon level of effort. If t_1 is the probability of a Type 1 error, then $r=1-t_1$. Similarly, in a Type 2 error, the Principal trusts an Agent that will in fact shirk. If t_2 is the probability of a Type 2 error, then $w=t_2$.

[22] One way to interpret TL is that it represents the risk the Principal faces when trusting the Agent, where the risk is affected by the probability that the Agent is trustworthy as well as the ratio of losses to gains from trusting.

Analysis and Implications

Figure 3 presents a graph of the TL (the minimal ratio of expected losses to expected gains from trusting) as π changes, holding constant the losses and gains. It shows that the TL increases as the probability that the Agent is trustworthy decreases, other things being equal. When the probability π that the Agent will honor the contract if trusted decreases (e.g., from π_1 to π_2), then the Principal will trust only if the RTC is increased (from RTC_1 to at least RTC_2). Conversely, an increase in π (e.g., from π_2 to π_1) will mean the Principal can more easily trust the Agent (the RTC can be relaxed). Note that in the extreme case in which $\pi = 1$, suggesting that the Principal knows with certainty that the Agent will honor the Principal's trust, then the Principal will always trust, even for very low RTC. However, when $\pi < 1$, that is, when trustworthiness is expected with probability but not with certainty – that is, when the Principal expects that the Agent might exploit the Principal's trust – then the Principal will trust only when $RTC \geq TL$. To be clear, the Principal can reliably trust when the RTC is in the "Trust" region of Figure 3 (i.e., when the RTC is equal to or exceeding the TL for a given probability of trustworthiness).

Moreover, an increase in the TL (e.g., from TL_1 to TL_2) will require an increase in the RTC (from RTC_1 to at least RTC_2) before the Principal would want to trust. For concreteness, suppose initially that $RTC_1 = TL_1$ (e.g., point A in Figure 3). An increase in the TL means that the Principal must feel more reliable in trusting than monitoring before she is willing to trust. That is, RTC_1, which is a function of the Principal's assessment of the relative temptation the Agent faces to exploit the Principal's trust, given that the Agent's probability of being trustworthy is π_1, must increase to at least RTC_2 (e.g., to point C) before the Principal will trust the Agent. If the RTC_1 does not increase, then the increase in TL will reduce the willingness of the Principal to trust – because the minimum trusting limit for reliably trusting is not

---

Generally, the risk from trusting increases as the probability of the Agent behaving honestly decreases, the loss from misplaced trust increases, or the gains from correctly trusting increases.

reached – thus leading the Principal to place a greater reliance on costly monitoring of the Agent. Conversely, a reduction in the TL means that the Principal can meet the reliability condition more easily (e.g., the RTC can decline), thus resulting in relatively more trust by the Principal, even when the Agent is expected to be less than fully trustworthy (e.g., movement of π_1 to π_2).

Generally, the trusting limit (TL) will decrease when the losses from misplaced trust, *L*, decrease, the gains from correctly trusting, *G*, increase, or the probability that an Agent is trustworthy, π, increases. Specifically, the TL will decrease when (1) the *a priori* probability increases that the Agent is trustworthy, (2) the probability decreases that the Agent, when monitored, will complete the contract, (3) the contracted action increases, (4) the wage payment decreases, and (5) the cost of monitoring increases.[23] When the TL decreases, the Principal can more reliably trust the Agent, given the Principal's assessment of the relative temptation the Agent has to shirk when trusted. However, when the TL increases, the Principal will likely monitor the Agent unless she can increase her reliability of trusting (the RTC). This could happen if the Principal perceives that the temptation of the Agent to shirk is reduced because, for example, the Agent credibly signals or Principal believes that the Agent has a higher quality of character than previously anticipated. Absent the increase in the RTC, the Principal will adopt a costly policy of monitoring the Agent as a result of the increase in the TL.

Furthermore, as seen in Figure 4, in cases in which the reliability condition is not satisfied (e.g., $RTC_1<TL_1$ for Agent with π_1), the Principal will not trust the Agent, even though the anticipated probability that the Agent behaves honestly exceeds π^* (e.g., $\pi_1 > \pi^*$)(see area M in Figure 4). In other cases in which the reliability condition is satisfied (e.g., $RTC_2>TL_2$ for Agent with π_2), some agents with $\pi < \pi^*$ will be trusted (see area T). For example, consider the following illustration that also expands the scope of the model. Suppose that the Principal can select an Agent from one of two pools of potential agents and that in each pool the average ability of workers is the same. The two pools of agents differ,

however, in the distribution of *H*-types.[24] Specifically, suppose that in the first pool the average distribution of *H*-types is $\overline{H}_1$ and in the second pool the average distribution of *H*-types is $\overline{H}_2$, where $\overline{H}_1 > \overline{H}_2$. Accordingly, the Principal expects that the probability that the Agent selected from the first pool will behave honestly is π_1 and that the probability the Agent from the second pool completes the agreed-upon contract is π_2. Because $\overline{H}_1 > \overline{H}_2$ we know that $\pi_1 > \pi_2$. Assume further that $\pi_1 > \pi^* > \pi_2$ so that, *a priori*, the Principal is inclined to trust the Agent selected from the first pool but monitor the Agent from the second pool. Now, suppose that after the Agent from each pool is selected the Principal learns that the Agent from the first pool is a cheat but the Agent from the second pool has high moral standards, so that the RTC for the first Agent is RTC_1 (where $RTC_1<TL_1$ for the first Agent with π_1 in Figure 4) but the RTC for the second Agent is RTC_2 (where $RTC_2>TL_2$ for the second Agent with π_2). Then, the Principal will trust the second Agent but monitor the first.

This analysis of trusting suggests the following implications. First, a principal may accept the fact that trusting could result in losses but choose to trust nonetheless because she perceives the relative temptation of an agent to exploit trust is small (either because the personal gain from exploitation is small or the quality of the agent's character is large). For example, managers may "trust" their employees not to pilfer office supplies in that managers take few precautions to monitor employee activity around office supplies, but they also accept that many employees will steal. The reason that managers trust in this circumstance is that the relative temptation felt by employees to pilfer is considered small (the gain to employees is pad of paper or a pen), so that the RTC is large (the low expected relative temptation means $\mathbf{X}$ is low, thus increasing the $\frac{r(\mathbf{X})}{w(\mathbf{X})}$ ratio or RTC), while the relative losses to managers are small, considering how costly it might be for them to establish effective monitoring of office supplies, so that

---

[23] See Appendix II for the corresponding propositions and proofs. Also, to be clear, condition (1) is manifested as a movement down along the TL curve graphed in Figure 3, while conditions (2) through (5) are manifested as downward shift of the TL curve.

[24] That is, the agents in each pool are assumed to have identical utility functions, except for their values of *H*.

the TL is relatively low. With the RTC relatively large in comparison with the TL, managers will be more inclined to trust than monitor in this circumstance. If, however, the total losses to employee theft are considered to be large, manifested as an increase in the TL, managers may take measures to monitor or otherwise control employee behavior around office supplies.

Second, even with relatively large losses from misplaced trust, manifested as large TL, a principal may not necessarily trust less. When the losses to misplaced trust are large, the question of whether a principal should trust depends on how reliable the principal is in the behavior of the agent. Thus, an agent with a reputation for trustworthiness will likely be trusted more often than an agent without such a reputation (Dasgupta, 1988, Kreps, 1990), though reputations alone are neither necessary nor sufficient to ensure trusting by a principal. Any factor that makes a principal's perception or assessment of the expected behavior of an agent more reliable will be important in affecting how likely the principal trusts the agent. The key is the principal's assessment of the relative temptation an agent faces to exploit trust rather than the absolute presence of incentives for shirking. For instance, the more certain the principal is in the personal characteristics of the agent, the more trust the principal can place in him, even when the potential loss of such trust is large. Empirically, this suggests that a principal would trust an agent who is known by the principal or who is close to the individual in characteristics. For example, Alesina and La Ferrara (2000) show that trust is reduced in more heterogeneous communities, and Glaeser et al (2000) show that trust in increased when individuals are close socially. Shared common experiences and group identity are also related to trust formation (see Dawes and Thaler, 1988). Communication is one element of a shared common experience, and empirically it is an important factor affecting the likelihood that players in a PD environment cooperate (Valley, Moag, and Bacerman, 1998; Tullock, 1999; Cohen and James, 2001; see also Dawes and Thaler, 1988). However, when the potential losses to misplaced trust are small relative to the potential gains from correctly trusting, then the principal may be more willing to trust the agent, even if the principal is unfamiliar with the agent's personal characteristics. This mirrors the analysis of Coleman (1990), who argues that some people may be "slow to trust a friend but quick to trust a confidence man" (p. 104) – the tolerance limit for trusting a friend, in which confidences and personal

weaknesses are shared, may exceed the tolerance limit the potential loss associated with a "stranger" on the street who offers "a deal you cannot refuse."

Third, although a principal's decision to trust or rely on costly monitoring depends on a variety of factors, not all of these factors will be perfectly known. Thus, a principal will rely on perceptions and guesswork in addition to economic and other forms of computational analysis in order to determine whether or not to trust, particularly early in the relationship between principal and agent when specific information on agent qualities is not known. The basic issue is one in which the principal becomes sensitized to appropriate or relevant information. Whether or not a principal trusts an agent depends not only on the principal's perceptions regarding how trustworthy she believes the agent to be, but also on environmental factors, as well as the specific conditions reflected in the job contract established between the principal and agent, such as wages paid and tasks required. The better the information acquired by a principal regarding the relative temptation of an agent to shirk when trusted, which is affected by environmental incentives as well as the personal characteristics of the agent, the better able the principal will be in assessing the merits of trusting relative to alternatives to trusting. Thus, principals face a tradeoff in the cost of acquiring information about the personal qualities of their agents and the cost of monitoring.

Fourth, when it is costly for a principal to assess the individual characteristics of potential agents there may be advantages to drawing on knowledge of group rather than individual characteristics because doing so reduces the minimum reliability necessary to induce trust (and hence avoid costly monitoring of agent behavior). For instance, if agents can be selected from one of two populations of equally qualified workers, but the principal knows that one population has a higher average probability of trustworthiness than the other (e.g., a higher average *H*-type for workers), then selecting agents out of that population means the reliability condition can be relaxed (a higher expected *H*-type, *a priori*, means the probability of trustworthiness is higher, thus moving the principal down the TL curve so that a lower RTC is required for trusting). This explains why, for instance, some managers hire workers from their alma mater rather than from schools they are not familiar with because it lowers their dependence on costly assessments of

individual agent characteristics. This also explains in part why Gorski (1993) finds that political leaders in Holland and Prussia selected bureaucratic workers on the basis of religious affiliation in order to save on administrative or monitoring costs, and why James (1999a, 1999b) finds that family firms utilize family members as employees to reduce agency costs.

Finally, in principal-agent relationships involving more than one agent, agents of like honesty may be trusted to different degrees – some may be trusted, and some may be monitored. Even under identical environmental conditions, identical agents (i.e., those with identical abilities and personal qualities) may not be equally trusted. What determines whether or not the principal trusts depends on the tolerance the principal has for losses associated with misplaced trust in an agent relative to potential gains from correctly trusting. This ratio is affected, in part, by the specific contracting requirements of the principal.

Conclusions

When will a principal trust an agent rather than rely on costly monitoring of the agent's performance? There are two general answers to this question. The first is obvious and occurs when the principal simply knows that the agent has an incentive to be trustworthy (Hardin, 1993; see also James, 2002). The second, however, arises when the possibility exists that the agent will exploit the principal's trust. As this paper has shown, a principal will trust when she believes she can reliably do so. In particular, the reliability of trusting is improved when the principal perceives that the relative temptation of the agent to exploit the principal's trust is low – either because the incentive to shirk is low or because the principal is confident about the personal characteristics of the agent. The reliability of trusting is also improved when the *a priori* probability of the agent being trustworthy increases (which could occur if the agent is selected from a population with a known "high" average level of trustworthiness relative to some other population of possible agents), as well as when the losses from misplaced trust relative to the gains from correctly trusting declines, as argued by Coleman (1990).

Although this model advances the technical analysis of trust, more effort needs to be directed at determining how trust and trustworthiness are developed and utilized in both economic and non-economic contexts. First, are environmental factors or personal characteristics more important in promoting trust? The answer to this question is likely to be empirical. Second, will the development of incentives to foster trust erode trust? There is some evidence that extrinsic incentives "crowd out" intrinsic incentives (Frey and Oberholzer-Gee, 1997; Frey and Jegen, 1999), which suggest that attempts to foster trust via incentives might ultimately undermine trust. These are important questions that should be explored more carefully in order to expand the economic understanding of trust.

Appendix I

Proof of $\frac{\partial \pi}{\partial a^*} < 0$: We first show that $\frac{\partial \tilde{H}}{\partial a^*} > 0$. Differentiating $\tilde{H}$ with respect to a^*, we obtain

$$\frac{\partial \tilde{H}}{\partial a^*} = \frac{\frac{d\,\mathrm{V}}{da^*}(a^* - a) - \mathrm{V}(a^*) + \mathrm{V}(a)}{(a^* - a)^2}.$$

To show that the numerator is positive, take a first order Taylor's series expansion of V(.) around a^*, which is

$$\mathrm{V}(a^*) = \mathrm{V}(a) + \frac{d\,\mathrm{V}}{da}(a^* - a).$$

Since $a^* > a$ and $\mathrm{V}(a)$ is assumed to be strictly increasing and convex, we know that $\frac{d\,\mathrm{V}}{da^*} > \frac{d\,\mathrm{V}}{da}$ by definition of convexity. Thus,

$$\mathrm{V}(a^*) < \mathrm{V}(a) + \frac{d\,\mathrm{V}}{da^*}(a^* - a).$$

Rearranging terms, we obtain

$$\frac{d\,\mathrm{V}}{da^*}(a^* - a) - \mathrm{V}(a^*) + \mathrm{V}(a) > 0.$$

Hence, $\frac{\partial \tilde{H}}{\partial a^*} > 0$. Since $\tilde{H}$ increases as the required action, a^*, increases, $\pi(H \geq \tilde{H})$ decreases. That is, $\frac{\partial \pi}{\partial a^*} < 0$. □

The proof of $\frac{\partial \phi}{\partial a^*} < 0$ is analogous.

Proof of $\frac{\partial \phi}{\partial w} > 0$: We need to show that $\frac{\partial \hat{H}}{\partial w} < 0$. Differentiating $\hat{H}$ with respect to w, we obtain

$$\frac{\partial \hat{H}}{\partial w} = \frac{-p}{a^* - a} < 0 .$$

Since $\hat{H}$ decreases as the wage payment, w, increases, then $\phi(H \geq \hat{H})$ increases. That is, $\frac{\partial \phi}{\partial w} > 0$. □

Proof of $\frac{\partial \phi}{\partial p} > 0$: First show that $\frac{\partial \hat{H}}{\partial p} < 0$ by differentiating $\hat{H}$ with respect to *p* to obtain

$$\frac{\partial \hat{H}}{\partial p} = \frac{-w}{a^* - a} < 0 .$$

Since $\hat{H}$ decreases as the probability a shirking Agent is detected increases, then $\phi(H \geq \hat{H})$ increases.

That is, $\frac{\partial \phi}{\partial p} > 0$. □

Appendix II

Proposition 1. The Trusting Limit (TL) will decrease if the probability increases that a trusted Agent honors the contract.

Proof: We show this by demonstrating that $\frac{\partial \mathrm{TL}}{\partial \pi} < 0$. Differentiating TL, we obtain

$$\frac{\partial \mathrm{TL}}{\partial \pi} = -\frac{L}{G\pi^2}.$$

The numerator is the loss, *L*, from misplaced trust, which by assumption is positive. Because the denominator is also positive, the expression is negative due to the preceding minus sign. Therefore, $\frac{\partial \mathrm{TL}}{\partial \pi} < 0$. □

Proposition 2. The Trusting Limit (TL) will decrease if the probability decreases that a monitored Agent honors the contract.

Proof: We show this by demonstrating that $\frac{\partial \mathrm{TL}}{\partial \phi} > 0$. Differentiating TL, we obtain

$$\frac{\partial \mathrm{TL}}{\partial \phi} = \frac{(1-\pi)[R(a^*-a)+(Ra-w)p]}{G\pi}.$$

The numerator is positive because $a^* > a$ and $Ra - w > 0$ by assumption, and the denominator is also positive. Therefore, $\frac{\partial \mathrm{TL}}{\partial \phi} > 0$. □

Proposition 3. The Trusting Limit (TL) will decrease if the contracted action, a^*, increases.

Proof: We show this by demonstrating that $\frac{\partial \mathrm{TL}}{\partial a^*} < 0$. Differentiating TL with respect *a**, we obtain

$$\frac{\partial \mathrm{TL}}{\partial a^*} = \frac{GL\frac{\partial \pi}{\partial a^*} + G\pi(1-\pi)[R(a^*-a)+(Ra-w)p]\frac{\partial \phi}{\partial a^*} - R\pi(1-\pi)[c(p)+(Ra-w)(p+(1-p)\phi)]}{G^2\pi^2}.$$

Because $\frac{\partial \pi}{\partial a^*} < 0$ and $\frac{\partial \phi}{\partial a^*} < 0$, the entire numerator is negative. Therefore, $\frac{\partial \mathrm{TL}}{\partial a^*} < 0$. □

Proposition 4. The Trusting Limit (TL) will decrease if the wage payment made to the Agent decreases.

Proof: We show this by proving $\frac{\partial \mathrm{TL}}{\partial w} > 0$ for sufficiently small *c*. Differentiating the TL with respect *w*, we obtain

$$\frac{\partial \mathrm{TL}}{\partial w} = \frac{(1-\pi)[G(R(a^*-a)+(Ra-w)p)\frac{\partial \phi}{\partial w} + R(a^*-a)(p+(1-p)\phi) - c(p)]}{G^2\pi}.$$

Because $\frac{\partial \phi}{\partial w} > 0$, the numerator is positive when

$c(p) < G(R(a^*-a)+(Ra-w)p)\frac{\partial \phi}{\partial w} + R(a^*-a)(p+(1-p)\phi)$. Therefore, $\frac{\partial \mathrm{TL}}{\partial w} > 0$ for $c(p)$ not too large. The idea here is that as the wage payment increases, the Principal is better off monitoring rather than trusting as long as the cost of monitoring is not too large. □

Proposition 5. The Trusting Limit (TL) will decrease if the cost of verifying performance increases.

Proof: We show this by proving $\frac{\partial \mathrm{TL}}{\partial c} < 0$. Differentiating the TL with respect *c*, we obtain

$$\frac{\partial \mathrm{TL}}{\partial c} = -\frac{(1-\pi)}{G\pi},$$

which is clearly negative. □

Figures

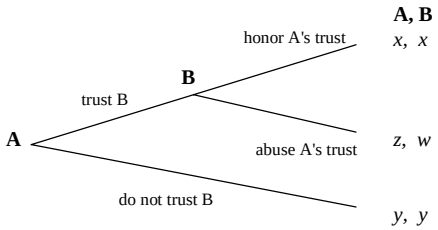

Figure 1. The "trust game" from Kreps (1990). The payoffs to player A are *x*, *z*, and *y*, and the payoffs to player B are *x*, *w*, and *y* where $z<y<x<w$.

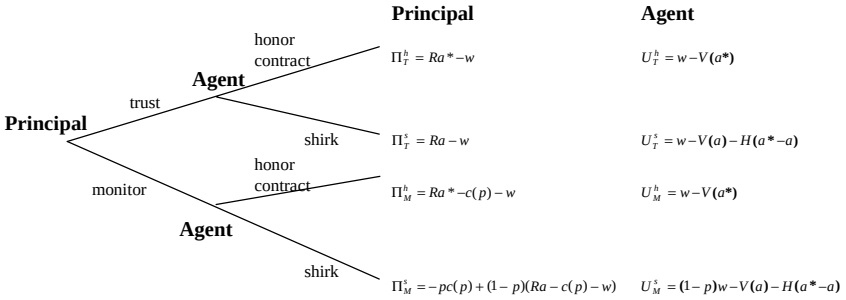

Figure 2. An extensive form representation of the Principal-Agent model. The Principal chooses either to trust or monitor the performance of an Agent, and the Agent chooses either to honor the contract or shirk.

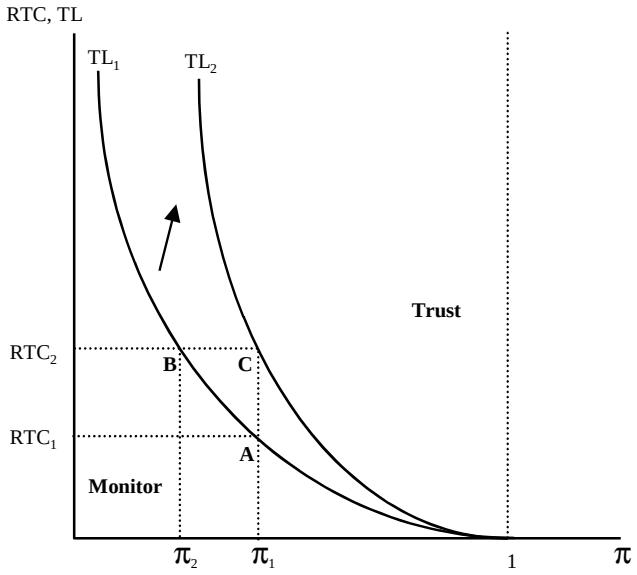

Figure 3. This figure shows how the Trusting Limit (TL) changes, holding constant the losses and gains from trusting, as π changes. Note that if π decreases, then the TL increases asymptotically. Accordingly, the principal must believe that trusting is more reliable (movement from point A to point B) before trusting. When the TL increases, then the principal must believe that trusting is more reliable (movement from point A to point C) before trusting.

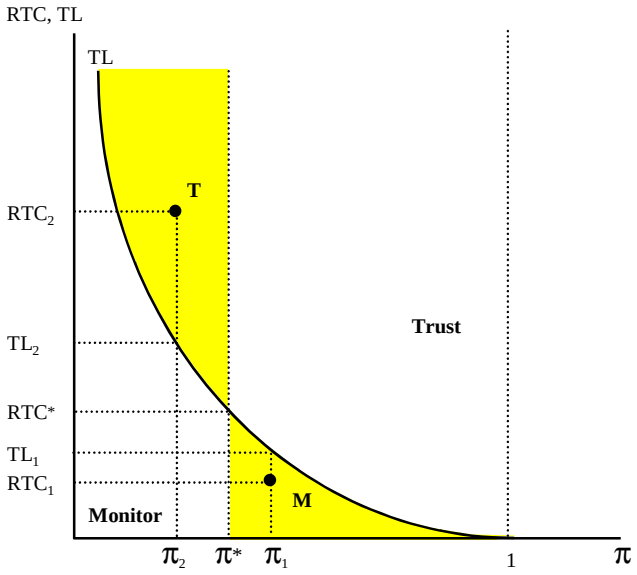

Figure 4. This figure shows how some agents with $\pi \geq \pi^*$ should not be trusted if RTC<TL in area M (e.g., $RTC_1<TL_1$ for agent with π_1), while other agents with $\pi < \pi^*$ should be trusted if RTC>TL in area T (e.g., $RTC_2>TL_2$ for agent with π_2), where π^* is the minimal probability of trustworthiness needed to induce trust by the principal independent of the RTC condition.

References

Alesina, Alberto, and Eliana La Ferrara, (2000) "The Determinants of Trust," working paper no. 7621, National Bureau of Economic Research.

Arrow, Kenneth J., (1973) Information and Economic Behavior, Stockholm: Federation of Swedish Industries.

Arrow, Kenneth J., (1974) The Limits of Organization, New York, NY: W.W. Norton.

Baier, Annette, (1986) "Trust and Antitrust," Ethics, 96(January), pp. 231-260.

Cohen, Jeffrey P., and Harvey S. James, Jr., (2001) "If Teaching Economics Discourages Cooperation, Can the Damage Be Undone?" working paper, Department of Economics, University of Hartford.

Coleman, James S., (1990) Foundations of Social Theory, Cambridge, MA: Harvard University Press.

Dasgupta, Partha, (1988) "Trust as a Commodity," in D. Gambetta (ed.) Trust: Making and Breaking Cooperative Relations, New York, NY: Basil Blackwell, pp. 49-72.

Dawes, Robyn M., and Richard H. Thaler, (1988) "Cooperation," Journal of Economic Perspectives, 2(3), pp. 187-197.

Frank, Robert H., Thomas D. Gilovich, and Dennis T. Regan, (1993) "Does Studying Economics Inhibit Cooperation?" Journal of Economic Perspectives, 7(2), pp. 159-171.

Frey, Bruno S., and Reto Jegen (1999) "Motivation Crowding Theory: A Survey of Empirical Evidence," working paper no. 26, Institute for Empirical Research in Economics, University of Zurich.

Frey, Bruno S., and Felix Oberholzer-Gee (1997) "The Cost of Price Incentives: An Empirical Analysis of Motivation Crowding-Out," American Economic Review, 87(4), pp. 746-755.

Glaeser, Edward, David I. Laibson, Jose A. Scheinkman, and Christine L. Soutter, (2000) "Measuring Trust," Quarterly Journal of Economics, 115(3), pp. 811-846.

Gorski, Philip, (1993) "The Protestant Ethic Revisited: Disciplinary Revolution and State Formation in Holland and Prussia," American Journal of Sociology, 99(2), pp. 265-316.

Güth, Werner, and Hartmau Kliemt, (1994) "Competition or Co-operation: On the Evolutionary Economics of Trust, Exploration and Moral Attitudes," Metroeconomica, 45(2), pp. 155-187.

Hardin, Russell, (1993) "The Street-level Epistemology of Trust," Politics & Society, 21(4), pp. 505-529.

Heiner, Ronald A., (1983) "The Origin of Predictable Behavior," American Economic Review, 73(4), pp. 560-595.

Holmstrom, Bengt, (1982) "Moral Hazard in Teams," Bell Journal of Economics, 13, pp. 324-340.

Holmstrom, Bengt, and Paul Milgrom, (1991) "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design," Journal of Law, Economics, and Organization, 7(spring), pp. 24-52.

Huang, Peter H., and Ho-Mou Wu, (1994) "More Order without More Law: A Theory of Social Norms and Organizational Cultures," Journal of Law, Economics, and Organization, 10(2), pp. 390-406.

Huck, Steffen, (1998) "Trust, Treason, and Trials: An Example of how the Evolution of Preferences Can Be Driven by Legal Institutions," Journal of Law, Economics, and Organization, 14(1), pp. 44-60.

James, Harvey S., Jr., (1999a) "What Can the Family Contribute to Business? Examining Contractual Relationships," Family Business Review, 12(1), pp. 61-72.

James, Harvey S., Jr., (1999b) "Owner as Manager, Extended Horizons and the Family Firm," International Journal of the Economics of Business, 6(1), pp. 41-55.

James, Harvey S., Jr., (2000) "Separating Contract from Governance," Managerial and Decision Economics, 21, pp. 47-61.

James, Harvey S., Jr., (2002) "The Trust Paradox: A Survey of Economic Inquiries Into the Nature of Trust and Trustworthiness," Journal of Economic Behavior and Organization, 47(3), forthcoming.

Jensen, Michael C., and William H. Meckling, (1976) "Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure," Journal of Financial Economics, 3, pp. 303-360.

Kay, Neil M., (1996), "The Economics of Trust," International Journal of the Economics of Business, 3(2), pp. 249-261.

Koford, Kenneth, (2000) "Monitoring in Organizations When Agents Vary in Ethical Standards," paper presented at the Eastern Economic Association meetings, Washington.

Kreps, David M., (1990) "Corporate Culture and Economic Theory," in J. Alt and K. Shepsle (eds.), Perspectives in Positive Political Economy, New York, NY: Cambridge University Press.

Lahno, Bernd Lahno, (1995) "Trust and Strategic Rationality," Rationality and Society, 7(4), pp. 442-464.

Leibenstein, Harvey, (1987) "On Some Economic Aspects of a Fragile Input: Trust," in G.R. Feiwel (ed.), Arrow and the Foundations of the Theory of Economic Policy, New York, NY: New York University Press, pp. 600-612.

Levi, Margaret, (2000) "When Good Defenses Make Good Neighbors: A Transaction Cost Approach to Trust, the Absence of Trust and Distrust," in C. Menard (ed.), Institutions, Contracts, and Organizations: Perspectives from New Institutional Economics, Northampton, MA: Edward Elgar Publishing, pp. 137-157.

North, Douglass C., (1990) Institutions, Institutional Change and Economic Performance, New York, NY: Cambridge University Press.

Quigley-Fernandez, B., F.S. Malkis, and J.T. Tedeschi, (1985) "Effects of First Impressions and Reliability of Promises on Trust and Cooperation," British Journal of Social Psychology, 24, pp. 29-36.

Snijders, Chris, and Gideon Keren, (1999) "Determinants of Trust," in D.V. Budescu, I. Erev, and R. Zwickm (eds.), Games and Human Behavior: Essays in Honor of Amnon Rapaport, Mahwah, NJ: Lawrence Erlbaum Associates, pp. 355-385.

Tullock, Gordon, (1999) "Non-prisoner's Dilemma," Journal of Economic Behavior and Organization, 39, pp. 455-458.

Valley, Kathleen L., Joseph Moag, and Max H. Bazerman, (1998) "'A Matter of Trust': Effects of Communication on the Efficiency and Distribution of Outcomes," Journal of Economic Behavior and Organization, 34, pp. 211-238.

Williamson, Oliver E., (1985) The Economic Institutions of Capitalism, New York, NY: The Free Press.

Williamson, Oliver E., (1993) "Calculativeness, Trust, and Economic Organization," Journal of Law and Economics, 36(April), pp. 453-486.