Hitotsubashi Journal of Economics 50 (2009), pp.205-224.

DEFINING TRUST USING EXPECTED UTILITY THEORY*

Kazuhiro Arai

Graduate School of Economics, Hitotsubashi University
Kunitachi, Tokyo 186-8601, Japan
kaz.arai@econ.hit-u.ac.jp

Accepted September 2009

Abstract

Trust has been discussed in many social sciences including economics, psychology, and sociology. However, there is no widely accepted definition of trust. In particular, there is no definition that can be used for economic analysis. This paper regards trust as expectation and defines it using expected utility theory together with concepts such as betrayal premium. In doing so, it rejects the widely accepted black-and-white view that (un)trustworthy people are always (un)trustworthy. This paper also discusses various determinants and properties of trust on the basis of the idea that trust is not simply a matter of intention.

Keywords: Definitions of Trust, Distrust Premium, Betrayal Premium, Properties of Trust, Expected Utility Theory
JEL Classification: D81, Z13.

I. *Introduction*

The concept of trust is becoming increasingly important in economics, but it has not yet been given a satisfactory definition that can be used for economic analysis. An important reason for this is that it is a concept that is quite remote from mainstream economics or neoclassical economics.[1] The purpose of this paper is to define trust using expected utility theory and discuss its basic determinants and properties.

Neoclassical economics does not explicitly discuss trust, which can be easily seen from the absence of this concept in it. Correspondingly, few economics textbooks for students mention trust. One of the basic reasons for this absence is that neoclassical economics assumes contract completeness. To put it more explicitly, it assumes that all transactions are performed under sufficiently detailed contracts and that no individual fails to comply with them.[2]

* A Grant-in-Aid for Scientific Research is gratefully acknowledged.

[1] In this paper, neoclassical economics stands for the general equilibrium theory developed by Arrow and Debreu (1954) and, in particular, by Debreu (1959).

[2] Neoclassical economics also implicitly assumes the completeness of the law. Namely, it assumes that the law is sufficiently detailed and all individuals comply with it. In the following, this paper will concentrate on contract completeness and will not refer to law completeness.

In essence, neoclassical economics assumes that all individuals are trustworthy in the sense that they perfectly comply with contracts and the law. Hence, it virtually assumes that all economic agents can and do perfectly trust other individuals as well. Indeed, since the neoclassical economic paradigm does not have the police or courts, it is theoretically necessary for it to assume that all individuals are perfectly trustworthy in the above sense.

In contrast, whether individuals are trustworthy or not is a serious problem in the real world because it has few complete contracts, which can be understood from the existence of room for behavioral discretion in many transaction cases. The reason for the prevalence of incomplete contracts is that it is prohibitively costly to make complete contracts because of the transaction costs. For instance, it is impossible in the case of a labor contract to specify how to work each minute of each day and what punishment to apply when the work is not done properly.

As mentioned above, an incomplete contract tends to generate discretionary behavior in the parties involved in it. This fact in turn generates interdependence or a game situation among them. This game is very likely to be the prisoner's dilemma game, and pursuit of self-interest will not lead to efficiency. Thus, it becomes very important whether or not one's transaction partner will behave ethically or as promised, namely trustworthily. This logic shows why trustworthiness (or trust) is economically important in transactions.

Trust and trustworthiness are important not only in human relations with certain formal contracts, but also in other relations without them. An example of the former is human relations within organizations and an example of the latter is neighborhood relations. Trust and trustworthiness are economically important because they tend to increase efficiency significantly by aiding cooperation among those involved and by reducing transaction costs such as monitoring costs. Thus, there are strong reasons why trust and trustworthiness are indispensable in many human relationships.

The structure of this paper is as follows: Section II discusses some typical definitions of trust that have been proposed in social sciences. Section III provides my basic definitions of trust using probability. Section IV elaborates them on the basis of expected utility theory. Section V extends the definitions in Section IV by introducing the concept of betrayal aversion. Section VI examines the performance of my definitions using specific utility functions and distribution functions for the trustee's behaviors. Section VII considers basic determinants of the degree of trust. Section VIII discusses some important properties of trust that have been pointed out by many trust researchers. Section IX inquires into the meaning of intention involved in trustworthiness. Section X concludes this paper.

II. *Typical Definitions of Trust*

There is virtually no definition of trust proposed by economists, which can be confirmed by searching for references in EconLit using 'trust' and 'definition' (or 'defining' or 'define') as keywords. Most definitions that are currently referred to in social sciences are those proposed by sociologists or psychologists. Before discussing my own definition, I would like to introduce and criticize some typical definitions proposed by sociologists and psychologists. They can be classified into two different categories: those that regard trust as behavior and those that regard it as expectation.

We discuss first some typical definitions that regard trust as behavior. Deutsch (1962) defines individual A's trust in individual B as A's behavior consisting of actions that (a) increase A's vulnerability (b) to B whose behavior is not under A's control (c) in a situation in which the penalty (disutility) A suffers if B abuses that vulnerability is greater than the benefit (utility) A gains if B does not.[3]

The following example promotes understanding of this definition. A parent (A) exhibits trusting behavior if A hires a baby-sitter (B) to go to see a movie. This action significantly increases A's vulnerability, since A cannot control B's behavior after leaving A's house. If B abuses that vulnerability, the penalty may be a tragedy that may adversely affect the rest of A's life; if B does not, the benefit will be the pleasure of seeing the movie.

Chiba (1997) proposes a similar definition. He claims that individual A trusts individual B (α) if A chooses a risky action (β) in a situation with essential uncertainty about B's behavior, (γ) anticipating that B will behave favorably towards A. Conditions (α) and (β) here nearly correspond to conditions (a) and (b) respectively in Deutsch's definition.

These two typical definitions have several shortcomings. First, in the real world, trust does not necessarily relate to the behavior of the truster toward the trustee or increase the truster's vulnerability to the trustee. For example, it is likely that A trusts judges even if A is not considering whether or not to use the courts. Another example is that A tells individual C that B is trustworthy or that A trusts B. In this case, A does not usually increase his vulnerability to B by saying so.

Thus, trust is a concept that is generally different from the behavior or vulnerability of the truster. It should be added that trust exists even if condition (c) in Deutsch's definition is not satisfied, though it is likely to be satisfied in many trust relations.

Chiba's definition does not specify what causes uncertainty about B's behavior. He probably considers the uncertainty A faces regarding B's intention, but the above definition can contain other uncertainty factors as well such as B's competence and the weather, which will affect B's behavior together with B's intention. This is also the case in Deutsch's definition.

Another problem with Chiba's definition is that the true meaning of favorable behavior is not clear because it has many different levels in the real world including minimally favorable behavior, fairly favorable behavior, and perfectly favorable behavior. Similarly, there are many levels in the penalty and the benefit of Deutsch's definition.

In addition, it is tautological to talk about 'a risky action' in (α) and 'uncertainty' in (β) in Chiba's definition. Moreover, since Chiba presupposes essential uncertainty about B's behavior, he faces the serious problem of being unable to define perfect trust (with no risk). Most trust studies seem to consider perfect trust as the ideal, but this definition cannot define such a state. Deutsch's definition also has the same problem.

Next, we discuss some typical definitions that regard trust as expectation. Sako (1992) gives the following definition: Trust is a state of mind, an expectation held by one trading partner about another, that the other behaves or responds in a predictable and mutually acceptable manner.

This definition does not escape serious shortcomings, either. For instance, it does not consider how the degree of trust changes in accordance with a variety of possible behaviors or responses of the other partner. If the expected behavior is only slightly less acceptable than a

[3] See also Zand (1972).

specific level, is the other trading partner completely untrustworthy? Since there are in general many different possible levels regarding the other partner's behavior, a definition of trust needs to take this fact into consideration.

On the other hand, Lazric and Lorenz (1998) emphasize the following three conditions as the basis for defining trust. (i) Trust is identified with an agent's beliefs rather than with his behavior or actions. (ii) Trust refers to beliefs about the likely behavior of another, or others, which matter for the truster's decision making. (iii) Trust pertains to situations where the complexity of the relationship, or the fact that it is marked by unanticipated contingencies, precludes having recourse to complete contingent contracts with third-party enforcement.

This idea also has some shortcomings. Although it deals with beliefs rather than behavior, those beliefs relate only to the truster's decision to interact with the trustee. I mentioned above that there are different types of trust in the real world. In addition, the above idea does not show how those beliefs are expressed.[4]

III. *Preliminary Definitions*

I would now like to discuss my own definitions of trust. I first proposed some definitions in Japanese in Arai (2000). Here, I would like to discuss them first and then propose extended versions. According to the above classification of definitions of trust, my definitions belong to the latter, i.e., I regard trust as expectation. In the following, I will discuss these definitions step by step from the simplest and most intuitive to the more general ones.

The simplest and most intuitive definition is the following.

Definition 1: Individual A trusts individual B if A expects B to keep B's promise or to comply with what is socially considered to be ethical (when B says nothing).

An important idea behind this definition is that trust needs to be defined in relation to a certain ethical criterion, which is either B's promise or what is socially considered to be ethical. Since human beings do not always make promises regarding their future behavior, what is socially considered to be ethical becomes the ethical criterion when no promise is made. This definition clearly shows that trust is regarded as a kind of expectation. It should be noted that this expectation may be purely subjective, namely, that even if A trusts B, individual C may not trust B. Moreover, it is likely that A trusts B in one respect but does not in another.

It can happen that some people make unethical promises like those in gangs or cliques within organizations. If the promise in the above definition is unethical, trust can be established even about unethical matters. If one does not want this feature to arise in a definition of trust, one can distinguish between ethical and unethical trust. Of course, ethics can vary across societies, so ethical trust in society X may be unethical in society Y. Less importantly, it can happen even in the same society or group that ethical judgments differ among different members.[5]

---

[4] See also Ring and Van de Ven (1992), Barney and Hansen (1994), and Zaheer et al. (1998). They treat trust as expectation as well and share similar shortcomings.

[5] It is also likely to happen within gangs or cliques that even if their members behave unethically without making unethical promises, trust can be established. This is because these groups have socially unethical norms. This kind of

There is a defect in Definition 1. Do we have to say that A does not trust B if there is 0.1% probability that B does not keep his promise? This defect can be overcome simply by introducing probability as the following second definition.

Definition 2: Individual A trusts individual B if A believes with a high probability that B will keep B's promise or comply with what is socially considered to be ethical (when B says nothing).

To be rigorous, 'trusts' in the above definition should be replaced by 'highly trusts' in correspondence with 'a high probability', but since 'trusts' implies 'highly trusts' in most daily conversations, this definition has followed the 'daily life' meaning. For the same reason as above, this probability can be purely subjective.

This definition shows that trust is not a matter of black and white, but a matter of degree. Many studies regard trust as a black-and-white problem: They consider either that A trusts B completely or that A does not trust B at all. Similarly, they assume either that B is completely trustworthy or that B is completely untrustworthy. In fact, the researchers whose definitions of trust have been discussed above have this black-and-white view of trust implicitly because they do not use probability in their definitions.

It should be added that Definition 2 is consistent with the definition of trust used in engineering. More specifically, trust is defined in engineering as the probability with which the item in question performs the work required in a given condition during a specified period.

An example will clarify the meaning of Definition 2. When A makes a decision as to whether to lend money to B, A's degree of trust in B about this particular matter is expressed by the probability with which A believes that B will repay A the debt (with interest). This example suggests that the degree of trust depends on a variety of factors including the amount of money to be lent. A may trust B when A is considering lending one thousand US dollars, but may not in the case of a hundred thousand dollars. This paper considers important determinants of trust in Section VII.

IV. *Definitions Using Expected Utility Theory*

Definition 2 is sufficiently general to be applied to many cases, but it is not fully general because it may happen in the above example that B returns half or two thirds of the amount B is obliged to return. In other words, the number of B's possible actions is more than two and the above-mentioned ethical criterion is partially satisfied in some of them.

This defect can be overcome by utilizing expected utility theory. In order to show it, suppose A is making a decision as to whether to lend money to individual B as above. A believes that B will pay back s_i with probability p_i, where $0 \leq p_i \leq 1$, $\sum_{i=0}^{n} p_i = 1$, $s_0 > s_1 > \cdots > s_k > s_{k+1} > \cdots > s_n = 0$, and s_0 is the full amount to be returned including interest. Let u be A's von Neumann-Morgenstein utility function and let $w + s_i$ be A's wealth when B returns s_i. Then, A's expected utility of lending the money is expressed as

trust should also be classified as unethical trust.

$$\sum_{i=0}^{n} p_i u(w+s_i). \tag{1}$$

An idea for a new definition is to use this expected utility level in relation to the upper limit of perfect trust and the lower limit of no trust. The upper limit stands for the case in which A believes that B will return s_0 with certainty and A's (expected) utility becomes equal to $u(w+s_0)$. On the other hand, the lower limit stands for the case in which A believes that B will return s_n or nothing with certainty and A's (expected) utility becomes equal to $u(w+s_n)$.

This idea will generate the degree of A's trust in B, which can be expressed by measuring how close the expected utility in expression (1) is to the upper limit. In order to compute it, let us introduce the following expression:

$$tu(w+s_0)+(1-t)u(w+s_n), \tag{2}$$

where $0 \leq t \leq 1$. This expression can be interpreted as A's expected utility in the case where he expects that B will return the full amount with probability t and nothing with probability $1-t$. According to Definition 2, the value of t measures A's degree of trust in B when B has only those two options.

Equating expression (1) with expression (2) and solving the equation for t generates the following measure of the degree of A's trust in B:

$$t_1 = \frac{\sum_{i=0}^{n} p_i u(w+s_i) - u(w+s_n)}{u(w+s_0) - u(w+s_n)}. \tag{3}$$

Obviously, this measure is invariant with respect to any affine transformation of u. Hence, this degree of trust is unique under B's same preferences. It is clear that t_1 in expression (3) satisfies condition $0 \leq t_1 \leq 1$. The case of $t_1 = 1$ indicates a situation where A perfectly trusts B, while the case of $t_1 = 0$ indicates a situation where A does not trust B at all.

Summing up, the above discussion has produced the following third definition of trust.

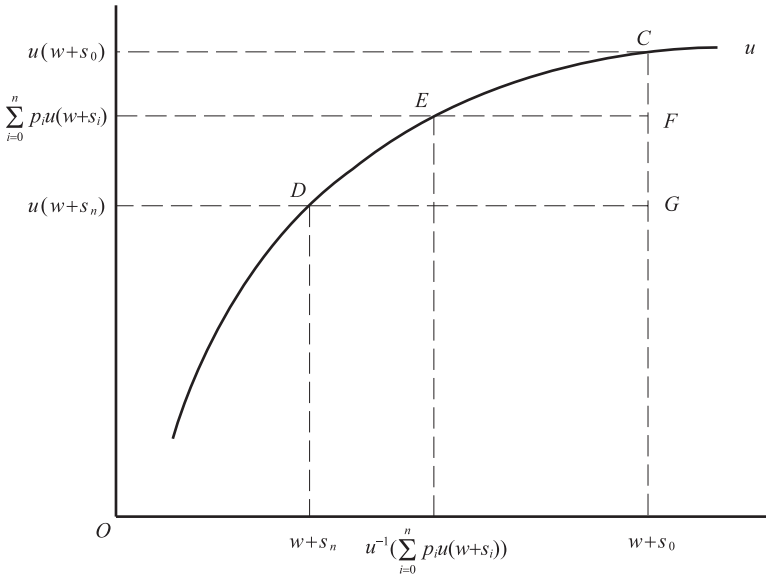
Definition 3: Suppose A is making a decision as to whether to interact with B. A believes that his wealth will become $w+s_i$ with probability p_i in accordance with action i B chooses against A from $n+1$ possible actions ($i=0, 1, 2, \cdots, n$), where $w+s_0$ is A's wealth when B perfectly keeps B's promise or perfectly complies with what is socially considered to be ethical, s_n is A's wealth when B behaves most poorly to A, and $w+s_k > w+s_{k+1}$. Under these circumstances, A's degree of trust can be expressed by expression (3).

Figure 1 illustrates the meaning of this third definition of trust. The denominator on the right-hand side in expression (3) equals distance CG in the figure, while the numerator equals distance FG. Thus, the following holds:

$$t_1 = \frac{FG}{CG}. \tag{4}$$

An observation of Figure 1 suggests that there can be another definition of trust or another way to express A's degree of trust in B. In order to see this, let us consider the meaning of distance EF. It can be interpreted as the loss A expects when he interacts with B, because A believes that B is untrustworthy to that extent. This distance can be called A's *distrust premium*

FIG 1



in the particular circumstances under consideration. Then, distance *DG* can be interpreted as the maximal possible distrust premium.

According to these interpretations, the following degree of trust can be defined:

$$t_2 = 1 - \frac{EF}{DG}. \tag{5}$$

Using the utility function, expression (5) can be rewritten as

$$t_2 = 1 - \frac{w + s_0 - u^{-1}\left(\sum_{i=0}^{n} p_i u(w+s_i)\right)}{s_0 - s_n}. \tag{6}$$

This gives the following fourth definition of trust.

Definition 4: Under the same circumstances as in Definition 3, A's trust in B can be expressed as expression (6).

The idea for Definition 4 is that the degree of trust is measured by how small the distrust premium is relative to the maximal level of the distrust premium. Note that the idea for the definition of distrust premium is different from that for risk (insurance) premium. Although the latter is defined as the distance between the expected return and the certainty equivalent, the former is defined as the distance between the ethical criterion ($w + s_0$) and the certainty equivalent $u^{-1}\left(\sum_{i=0}^{n} p_i u(w+s_i)\right)$. I think this is quite natural because the degree of trust needs to be measured in relation to a certain ethical criterion as mentioned above. It should be added

that the degree of trust expressed in expression (6) is also invariant with respect to any affine transformation of u.

Here, I would like to propose a slightly more general definition of trust. In the above cases of this section, B's choice of action i uniquely determines A's wealth $w+s_i$ or welfare $u(w+s_i)$, but this does not hold in some cases in the real world. For instance, even if a doctor chooses a specific treatment, the patient's welfare can also be affected by other factors such as B's physical condition, B's genetic or acquired factors, the condition of the medical equipment, the weather, the assistants' work, and so on. Similarly, even if a tourist guide has chosen a specific service, the welfare of the tourist who hires him also depends on the weather, traffic conditions, and crowdedness of the sightseeing spots, and so on.

These stochastic factors can be incorporated to generate a slightly more general definition of trust than the above. Here, I consider generalization of Definition 3 only. As above let p_i denote the probability with which A believes that B will choose action i ($0 \leq p_i \leq 1$). Under the circumstances considered here, A's welfare depends not only on B's action, but also on other stochastic factors. Let x_{ij} denote the state that will arise with probability π_{ij} when B chooses action i ($j=0, 1, 2, \cdots, m$). Number of states m can differ from action to action on the part of B, but we assume here that it is the same without loss of generality. In expression (3), x_{ij} is A's wealth with $x_{ij}=w+s_i$ for all j. A more general case is considered here.

As above, let action 0 be the best for A, action 1 be the second best, and so on. The following relations then hold:

$$\sum_{j=0}^{m} \pi_{0j} u(x_{0j}) \geq \sum_{j=0}^{m} \pi_{1j} u(x_{1j}) \geq \cdots \geq \sum_{j=0}^{m} \pi_{nj} u(x_{nj}). \tag{7}$$

In other words, the probability distribution corresponding to action k dominates that corresponding to action $k+1$. Using these conditions, A's degree of trust τ in B can be defined as

$$\tau = \frac{\sum_{j=1}^{m} \{ \sum_{i=0}^{n} p_i \pi_{ij} u(x_{ij}) - \pi_{nj} u(x_{nj}) \}}{\sum_{j=1}^{m} \{ \pi_{0j} u(x_{0j}) - \pi_{nj} u(x_{nj}) \}}. \tag{8}$$

This is an extension of expression (3). Hence, the following new definition can be proposed.

Definition 5: When there are risk factors other than B's action, A's trust in B can be expressed by expression (8).

V. *Definitions of Trust with Betrayal Aversion*

The definitions of trust discussed in the previous sections treated risks generated by human interactions in the same manner as those generated by asocial factors such as weather and earthquakes. In recent years, several researchers emphasize the differences between the two. They include Fehr and Schmidt (1999), Bolton and Ockenfels (2000), Bohnet and Zeckhauser (2004), Engelmann and Strobel (2004), Hong and Bohnet (2007), Bohnet et al. (2008), and

Fehr (2009).

These researchers distinguish the two types of risk by introducing the concept of betrayal aversion, which means that individuals are less willing to accept betrayal risks. Since risks inherent in trust intrinsically contain betrayal risks, they claim that those risks need to be treated differently from asocial risks.

Fehr (2009) says that people are more willing to take risk when facing a given probability of bad luck than to trust when facing an identical probability of being cheated. The idea behind the concept of betrayal aversion is the existence of special distaste for being a sucker or being exploited by untrustworthy partners. In many cases, betrayal aversion means that people have a dislike of non-reciprocated trust.

Bohnet et al. (2008) point to two sources of betrayal aversion. First, the trustee's decision determines not only the truster's payoff, but also the trustee's, in which the truster is highly interested. Second, elements beyond mere outcome-based preferences are likely to enter the utility function. Such elements include the psychological costs inherent in being betrayed.

Betrayal aversion can be incorporated into the trust definitions discussed above. In order to do so, we start with Definition 3 and extend it by introducing the concept of *betrayal premium*.

Assume that A's utility equals $u\ (w+s_i-b_i)$ when B chooses action i, where $b_i \geq 0$ denotes A's betrayal premium and $0=b_0 \leq b_1 \leq \cdots \leq b_n$. This assumption means that when A faces a trust problem, his welfare depends not only upon his pecuniary state, but also upon his psychological state generated by B's action. Note that the betrayal premium increases as the size of B's betrayal becomes larger, though the premium equals zero in the case where B perfectly keeps his promise or perfectly complies with what is socially considered to be ethical.

Using the concept of betrayal premium, Definition 3 can be extended to the following.

Definition 6: When individual A has betrayal premium $b_i \geq 0$ for individual B's action i under the same circumstances as in Definition 3 ($0=b_0 \leq b_1 \leq \cdots \leq b_n$), A's trust in B can be expressed as expression (9).

$$t_1=\frac{\sum_{i=0}^{n} p_i u(w+s_i-b_i)-u(w+s_n-b_n)}{u(w+s_0-b_0)-u(w+s_n-b_n)}. \tag{9}$$

As before, Figure 2 shows the graphical meaning of Definition 6. Because of the existence of betrayal aversion, the distrust premium here is larger than that in Figure 1. In particular, the maximal possible distrust premium where B chooses action n equals distance HK, which is larger than distrust premium DG by the amount of betrayal premium HL.

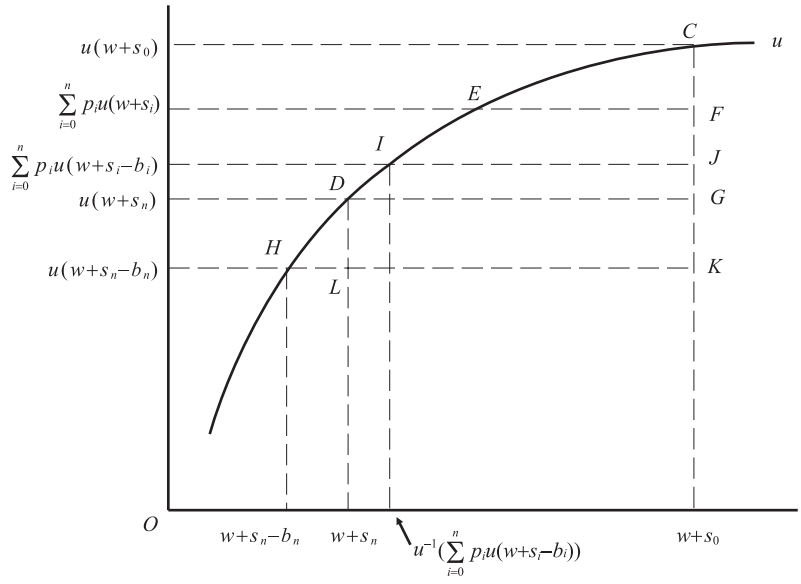
In Figure 2, A's degree of trust in B defined in Definition 6 can be expressed as the following ratio:

$$t_1=\frac{JK}{CK}. \tag{10}$$

Another new definition using the betrayal premium can also be made following the idea of Definition 4. Then, expression (5) needs to be replaced with

$$t_2=1-\frac{IJ}{HK}. \tag{11}$$

FIG 2



Using the utility function, expression (11) can be rewritten as

$$t_2=1-\frac{w+s_0-u^{-1}(\sum_{i=0}^{n} p_i u(w+s_i-b_i))}{s_0-s_n+b_n}. \tag{12}$$

Therefore, we have the following definition.

Definition 7: Under the same circumstances as in Definition 6, A's trust in B can be expressed as expression (12).

VI. *Numerical Examples of Degree of Trust*

This section provides several numerical examples of degree of trust using some of the definitions of trust given in the previous sections. We use Definitions 3, 4, 6, and 7 for the examples here, because they are the representative definitions of this paper. The primary purpose of considering the examples here is to examine whether these definitions have nice properties.

The examples use the following special utility functions:

$$u(x)=-e^{-\alpha x} \quad \alpha>0 \text{ and} \tag{13}$$

$$v(x)=-x^{1-r} \quad r>1, \tag{14}$$

where x denotes the amount of wealth of individual A or the truster. The utility function in

expression (13) exhibits the property of constant absolute risk aversion α in the Arrow-Pratt sense, and the larger the level of α, the larger the level of absolute risk aversion. On the other hand, the utility function in expression (14) exhibits the property of constant relative risk aversion r, and the larger the level of r, the larger the level of relative risk aversion. In the following, we consider the cases of $\alpha=1$, $\alpha=2$, and $\alpha=3$ for the utility function in expression (13), and $r=2$, $r=3$, and $r=4$ for the utility function in expression (14).

Let $n=2$, $w=8$, $s_0=2$, $s_1=1$, and $s_2=0$. Then, $w+s_0=10$, $w+s_1=9$, and $w+s_2=8$. The probability distribution of A's wealth when B is expected to pay back s_i with probability p_i is shown by the following notation ($i=0, 1, 2$):

$$d=\{w+s_0,\ w+s_1,\ w+s_2\colon p_0,\ p_1,\ p_2\}. \tag{15}$$

This means that A believes $w+s_i$ will occur with probability p_i.

For Definitions 3 and 4 we consider the following three probability distributions:

$$d_1=\{10,\ 9,\ 8\colon 0.8,\ 0.1,\ 0.1\}, \tag{16}$$

$$d_2=\{10,\ 9,\ 8\colon 0.7,\ 0.2,\ 0.1\}, \text{ and} \tag{17}$$

$$d_3=\{10,\ 9,\ 8\colon 0.6,\ 0.2,\ 0.2\}. \tag{18}$$

These example distributions imply that A trusts B more in the case of d_1 than in the case of d_2, and more in the case of d_2 than in the case of d_3.

Table 1 shows degrees of trust computed using the utility function in expression (13). It can be seen that the degree of trust is larger in the case of d_1 (d_2) than in the case of d_2 (d_3) for both t_1 and t_2 or for both Definitions 3 and 4 at each level of α. On the other hand, as A's degree of risk aversion increases, his degree of trust increases in the case of t_1 or Definition 3 and decreases in the case of t_2 or Definition 4. Hence, Definition 4 seems to have a better property as far as these numerical examples are concerned.

Next, we compute the degree of trust by introducing betrayal aversion. In order to see the effects of the magnitude of betrayal aversion, we consider two different cases.

In the first case, we assume that $b_0=0$, $b_1=0.5$, and $b_3=1$. Then, we have the following

TABLE 1. DEGREES OF TRUST

$u(x)=-e^{-\alpha x}$

		t_1	t_2
$\alpha=1$			
	d_1	0.8731	0.7031
	d_2	0.8462	0.6578
	d_3	0.7462	0.5182
$\alpha=2$			
	d_1	0.8881	0.5136
	d_2	0.8762	0.4917
	d_3	0.7762	0.3588
$\alpha=3$			
	d_1	0.8953	0.3725
	d_2	0.8905	0.3653
	d_3	0.7905	0.2590

TABLE 2. DEGREES OF TRUST WITH BETRAYAL AVERSION 1

$u(x) = -e^{-\alpha x}$

		t_1	t_2
$\alpha=1$			
	d_4	0.8818	0.6064
	d_5	0.8635	0.5726
	d_6	0.7635	0.4309
$\alpha=2$			
	d_4	0.8953	0.3725
	d_5	0.8905	0.3653
	d_6	0.7905	0.2590
$\alpha=3$			
	d_4	0.8989	0.2545
	d_5	0.8978	0.2533
	d_6	0.7978	0.1776

TABLE 3. DEGREES OF TRUST WITH BETRAYAL AVERSION 2

$u(x) = -e^{-\alpha x}$

		t_1	t_2
$\alpha=1$			
	d_7	0.8881	0.5136
	d_8	0.8762	0.4917
	d_9	0.7762	0.3588
$\alpha=2$			
	d_7	0.8982	0.2852
	d_8	0.8964	0.2830
	d_9	0.7964	0.1987
$\alpha=3$			
	d_7	0.8998	0.1917
	d_8	0.8995	0.1915
	d_9	0.7995	0.1339

distributions of A's real wealth that takes account of his betrayal premiums:

$$d_4 = \{10, 8.5, 7: 0.8, 0.1, 0.1\}, \tag{19}$$

$$d_5 = \{10, 8.5, 7: 0.7, 0.2, 0.1\}, \text{ and} \tag{20}$$

$$d_6 = \{10, 8.5, 7: 0.6, 0.2, 0.2\}. \tag{21}$$

In the second case, we assume that $b_0 = 0$, $b_1 = 1$, and $b_2 = 2$. The corresponding distribution of A's real wealth then becomes the following:

$$d_7 = \{10, 8, 6: 0.8, 0.1, 0.1\}, \tag{22}$$

$$d_8 = \{10, 8, 6: 0.7, 0.2, 0.1\}, \text{ and} \tag{23}$$

$$d_9 = \{10, 8, 6: 0.6, 0.2, 0.2\}. \tag{24}$$

Table 2 and Table 3 show the computation results for the first and second cases,

respectively. Although the degrees of trust for t_1 in Table 2 are larger than the corresponding values in Table 1, those for t_2 are smaller. Hence, t_2 or Definition 4 again behaves better as a definition of trust. A similar relationship exists between Tables 2 and 3.

The above degrees of trust were computed using the utility function in expression (13). The computation results for the utility function in expression (14) are shown in Tables 4, 5, and 6. They reveal properties quite similar to those observed in Tables 1, 2, and 3.

These observations suggest that degree of trust t_2 has many desirable properties. Although t_1 has the property of reducing the degree of trust for probability distributions with more distrust, its value increases when the degree of risk aversion increases or when the betrayal premium increases. Therefore, Definitions 4 and 7 seem to be the best definitions as far as the above numerical examples are concerned.

TABLE 4. DEGREES OF TRUST

$v(x) = -x^{1-r}$

		t_1	t_2
$r=2$			
	d_1	0.8556	0.8258
	d_2	0.8112	0.7746
	d_3	0.7112	0.6633
$r=3$			
	d_1	0.8578	0.8113
	d_2	0.8169	0.7608
	d_3	0.7173	0.6444
$r=4$			
	d_1	0.8605	0.7962
	d_2	0.8217	0.7451
	d_3	0.7220	0.6232

TABLE 5. DEGREES OF TRUST WITH BETRAYAL AVERSION 1

$v(x) = -x^{1-r}$

		t_1	t_2
$r=2$			
	d_4	0.8588	0.8098
	d_5	0.8175	0.7582
	d_6	0.7177	0.6402
$r=3$			
	d_4	0.8629	0.7852
	d_5	0.8260	0.7341
	d_5	0.7260	0.6072
$r=4$			
	d_4	0.8674	0.7574
	d_5	0.8059	0.6664
	d_6	0.7345	0.5729

TABLE 6. DEGREES OF TRUST WITH BETRAYAL AVERSION 2

$v(x) = -x^{1-r}$

		t_1	t_2
$r=2$			
	d_7	0.8624	0.7900
	d_8	0.8249	0.7387
	d_9	0.7250	0.6127
$r=3$			
	d_7	0.8684	0.7505
	d_8	0.8367	0.7009
	d_9	0.7367	0.5633
$r=4$			
	d_7	0.8737	0.7046
	d_8	0.8475	0.6585
	d_9	0.7475	0.5126

VII. *Determinants of Trust*

If trust is defined as in the previous sections, an interesting question that naturally arises is what the main factors are determining the degree of trust. This section considers this question mainly using the above hypothetical case in which A is making a decision as to whether to lend money to B. In this case, A's degree of trust in B depends above all upon the following five categories of factors: (I) the social environment, (II) B's characteristics, (III) the relationship between A and B, (IV) the characteristics of the object regarding which A trusts B, and (V) A's characteristics. Below, I would like to discuss in detail these factors one by one.

First, it may be obvious to many people that A's trust in B is affected by the social environment such as the culture, the legal system, and other characteristics of the society and/or organization to which A and B belong. A would not expect the money he has lent to be returned with a high probability if the culture does not attach much importance to defaulting on payment of debt. It is clear that the probability of repayment also depends on the legal system, i.e., the provisions of the law, how they are applied, the judicial system, the costs of using the courts, and so on. If the members of the society and/or organization to which A and B belong are individualistic and few righteous third persons (friends, relatives, colleagues, etc.) intervene in the dispute between A and B, B is expected to be less likely to try to repay the money, resulting in A's low trust in B.

Secondly, A's trust in B also depends upon B's characteristics such as his values, personality, earnings, economic conditions, competence, and so forth. It is obvious that these factors affect the probability with which B will return the money. Though the society's culture influences the values of its members, the extent of influence varies across individuals, that is, there is individual diversity in the extent to which the culture is internalized.

Thirdly, the social relationship between A and B obviously affects the probability with which B will return the money. If they are close friends, the probability must be high. If the relationship is expected to continue for a long period of time, the probability must also be high. The probability of repayment is higher in the case where the two individuals are neighbors than in the case where they live hundreds of kilometers apart. Similarly, the probability of

repayment is higher in the case where they belong to the same organization than in the case where they belong to different organizations.

Fourthly, the characteristics of the object regarding which A trusts or distrusts B are important determinants of A's degree of trust in B. The most important characteristic in the case of lending money is the amount of money to be returned. In general, the smaller the amount, the larger the probability of repayment, although the probability of repayment may be low for very small amounts such as a few dollars. When repayment is due may be another important characteristic. Important characteristics vary from case to case. In a different example where a boss is thinking about telling his subordinate to do a task, the difficulty of the task is a very important characteristic.

Fifthly, since the probabilities used to define trust are purely subjective as discussed above, A's characteristics affect his trust in B as well. For instance, A's family background strongly affects his expectations towards others. More generally, what experiences A has had in his life influences those probabilities. Older people might trust others less than younger people because the perceptiveness of the former is greater. In addition, as discussed in the previous sections, A's risk aversion and betrayal aversion affect his degree of trust.

VIII. *Basic Properties of Trust*

This section reexamines a few important properties of trust that have been pointed out by trust researchers and shows that many researchers have logically wrong ideas about trust concerning those properties. More specifically, this section first criticizes the widely (and implicitly) accepted view that society is made up of those who are trustworthy and those who are untrustworthy. It then criticizes the view that trust studies should focus on the intention involved in behavior by eliminating all other factors that generate desirable behavior such as competence and external pressures.

The first of these two views claims essentially that those who are trustworthy are always completely trustworthy and that those who are untrustworthy are always completely untrustworthy. This is a black-and-white view of trust. It seems that this same view is common to all researchers whose definitions of trust were discussed in Section II. Interestingly, this view will lead logically to the unwanted conclusion that studying trust is worthless. There are two reasons for this.

One reason is that if trustworthy people were always trustworthy and untrustworthy people were always untrustworthy, a low-cost experiment could be used to tell whether or not any particular individual is trustworthy. For instance, a prisoner's dilemma game could be used in an experiment to see if that person chooses the cooperative strategy.[6] If he chooses it, he will be trustworthy in any situation no matter how large his benefit from cheating will be. If this method is slightly fictitious, individual A could use an actual situation involving trust with a small amount of possible loss. If B behaves trustworthily in this instance, A can trust him perfectly in any trust situation no matter how large the possible loss will be. In this way, it

[6] In accordance with my definitions of trust, individual A may need to obtain from individual B a promise that B will be cooperative. However, this is not necessary in most cases because what is considered to be ethical in most societies is to behave cooperatively in a situation like the prisoner's dilemma game.

would be very easy and virtually costless to distinguish trustworthy from untrustworthy people.

The other reason is that it would be impossible to promote trust by any means if untrustworthy people were always untrustworthy. An important purpose of studying trust is to devise ways to promote trust. In fact, many people and organizations in the real world are trying to promote it by maintaining good human relations, exhibiting leadership, establishing institutions that foster it, and so on. However, the above view essentially regards all these efforts as useless.

If a low-cost experiment could be used to tell whether or not any particular individual is trustworthy, trust would actually become a matter of certainty, not uncertainty. Hence, the most important element of trust would disappear if the above black-and-white view were accepted. On the other hand, if it were impossible to promote trust by any means, trust would not be an object of economic analysis because it is of no use allocating resources or expending efforts to promote trust.

All these mean that studying trust would be worthless if the above black-and-white view of trust were accepted. This conclusion in turn reveals the importance of defining and analyzing trust using probability.

Next, we examine another basic property of trust. That is, we consider whether the elements of competence and social pressures can be eliminated from the trust concept to have only intention in it. Some trust researchers such as Barber (1983) and Yamagishi (1998) claim that lack of competence may prevent a trustworthy individual from behaving trustworthily. In other words, even if an individual behaves like an untrustworthy person, they claim, he may be trustworthy if his competence is low.

When B has borrowed money from A and promised to return it, B may fail to keep it simply because he is not competent enough to make money, even though he tries hard to return it. The above researchers regard B in this case as trustworthy because he has the intention of returning the money. They claim that only intention should be the object of analysis regarding trust. According to this view, a competent individual is not necessarily trustworthy even if he always keeps his promise.

The above researchers use similar logic to eliminate external pressures such as social or legal pressures from the trust concept. According to this logic, an individual who behaves well when he faces social or legal pressures against violation cannot be judged to be trustworthy because he may simply be avoiding sanctions by doing so. Sanctions may be implemented by some members of the community he belongs to, by some coworkers, by the law, or by someone else.

A special kind of social pressure arises when the trading partner has strong bargaining power. This holds, for example, in the relationship between a car assembler and a parts producer and that between a department store and a supplier. In this relationship the bargaining power on the part of the car assembler and the department store acts as social pressure because it can be used to terminate the relationship when a somewhat undesirable response is observed on the part of the parts producer or the supplier. According to the above view, a parts producer and a supplier cannot be said to be trustworthy even if they always deliver high-quality products on the appointed date, because they are under pressure.

What is common among these ideas is the view that analysis of trust should focus on intention by eliminating all other factors that generate desirable behavior. According to this view, trustworthy behavior should be chosen voluntarily without any external pressures. It may

look plausible at first sight, but it makes trust studies virtually impossible and excludes many interesting questions from analysis. Hence, I do not agree with this view. Below, I would like to expand on the reasons.

Firstly, there are many determinants of trust as discussed above and eliminating only competence and external pressures does not result in extracting intention. Eliminating them does not eliminate the influences of (a part of) the human relations between A and B, the culture of the society and/or the organization A and B belong to, and A's experiences, perceptiveness, and risk aversion.

For instance, whether A and B are in a long-term relationship affects trust and cooperative behavior as the theories of repeated games suggest, but it is independent of competence and external pressures. So is the effect of organizational cultural aspects such as smooth communication among coworkers. Game experiments by Dawes et al. (1977) and van de Kragt et al. (1983) demonstrate that pre-play communication significantly increases cooperation. Arai (1995) and Arai (2005) demonstrate that the experimenter's persuasion of players to cooperate also increases cooperation. Since pre-play communication and persuasion do not involve sanctions, they can be considered to be independent of external pressures.

Secondly, it is conceptually and technically impossible in almost all cases to eliminate competence and external pressures from the trust concept to extract only intention. For instance, what is the minimal income or wealth for a person who has borrowed twenty thousand US dollars and is regarded as competent enough to return the money? Most people in advanced countries can return that debt by cutting their food expenditure by one third for a few years without damaging their health (while actually improving it). Is such an act within or beyond their competence? No one seems to be able to answer this question. In fact, the above-mentioned researchers themselves are unlikely to be able to answer it. I think that they use the word 'competence' ambiguously and that they are unable to define it. If so, they cannot eliminate it theoretically from the trust concept. It is easy to use the word 'competence', but it is difficult to specify the conditions that generate competence. Many kinds of competence are beyond definition.

It should be added that a trustworthy individual does not borrow money that he is unable to return. Neither does he accept work he is unable to complete by the appointed time. Trustworthy people make promises and accept orders in accordance with their competence, so it is unnecessary to eliminate the competence factor from the trust concept.

IX. *The Essence of Intention*

It needs to be noted that intention is not necessarily independent of social or legal pressures. Ethics is internalized within human beings through socialization, which is nothing but a result of social and legal pressures. Many people might think that they would not commit homicide even if the law did not have punishments for it, but there is no doubt that the law promotes internalization of ethics. Moreover, there is no individual in the world who internalizes all ethics, so conscience works well only under social and legal pressures. Incidentally, those researchers who have the above-mentioned black-and-white view of trust need to believe that there are many people who have so completely internalized ethics that they behave ethically under any conditions in real society. This is too simplistic and unrealistic a

belief.

For the reasons given above, it is difficult to conceptualize trust that has eliminated social and legal pressures to extract only intention. What the above researchers regard as intention is very likely to be the result of psychological and cultural pressures through law and religion on the one hand and the pressures of the communities such as organizations and religious groups on the other. Many of the behaviors that are considered to be based on internal motivation are actually nothing but reactions to such invisible pressures. Indeed, even most behaviors that are considered to derive from free will are influenced by culture, although the degree of influence varies across individuals.

The view that deals with only intention in trust studies naturally has to accept the idea that behaviors consistent with self-interest are not trustworthy behaviors, because the above-mentioned behaviors whereby sanctions are avoided are equivalent to the pursuit of self-interest. As shown below, this idea also has a serious theoretical problem, which is another reason I disagree with this view.

At first sight, trust behaviors deriving from intention might seem quite different from those deriving from self-interest. There are many cases, however, in which it is difficult to distinguish between them. Indeed, if one tries to make a distinction, one is led to a strange conclusion.

The problem of reputation clarifies the point. In many cases, behaviors that generate good reputations are considered to be trustworthy behaviors. On the other hand, since individuals are fond of good reputations about themselves, behaviors that generate good reputations are consistent with self-interest. Therefore, if behaviors consistent with self-interest were not trustworthy behaviors, desirable behaviors that generate good reputations would not be trustworthy behaviors against the normal sense of human beings.

An interesting example is the following: Suppose that an individual has been behaving trustworthily, which has given him a good reputation, but that he himself has not heard of it. Suppose further that his reputation has become so enormous that one day he hears of it. Then, even if he behaves in the same way as before from that day on, he is conscious of his reputation. According to the above idea, his behaviors from that day on are not trustworthy behaviors because they are consistent with his self-interest. The exact same behavior is regarded as trustworthy when he does not know of his reputation but as untrustworthy when he knows of it. This is a very strange claim.

As another example, suppose that a large well-known corporation is providing conscientious care to its customers. It does not commit any injustice and actively provides its customers with all useful information. This kind of corporate behavior is not trustworthy behavior according to the above view, because it is likely to contribute to that corporation's profits. This is also a strange claim.

We have already mentioned that what is generally called trust cannot be determined only by intention, competence, or external pressures. However, the concept of 'conditions' proposed by Nooteboom (2002) is so general that it seems to include many determinants of trust. In contrast to the above-mentioned black-and-white view of trust, he claims that trust is determined by several conditions. This claim is closer to my view of trust, since those conditions can include the culture of the society that A and B belong to and (a part of) their social relations as conditions determining trust. However, even if this concept is introduced, there are still other determinants of trust such as A's experiences, perceptiveness, and risk aversion that I have pointed out above.

These considerations are also useful when undertaking international comparisons of trust. To see this, suppose that behavior X is considered to be desirable and equally widely observable in Societies 1, 2, and 3. Suppose further that members in Society 1 are punished by law if they do not exhibit X, that those in Society 2 cannot be promoted in their organizations if they do not exhibit X, and that those in Society 3 are frowned at by their religious group members if they do not exhibit X.

Those researchers who emphasize intention in trust studies must regard people in Society 3 as most trustworthy. However, these people are simply exhibiting X under the psychological and social pressures of religious groups. This is an attitude where psychological and social sanctions are feared, and it is similar to when legal or organizational sanctions are feared. Hence, members in Society 3 cannot be claimed to be more honorable or trustworthy than those in Societies 1 and 2.

All the above discussions suggest that it is impossible and improper to extract and analyze only intention in trust studies. As my definitions of trust describe, trust is nothing but a matter of the truster's expectations, which are determined by the many factors pointed out above.

X. *Conclusions*

There was previously no definition of trust that could be used for economic analysis. This paper has regarded trust as expectation or a subjective probability and defined it using expected utility theory together with concepts such as betrayal premium. In doing so, it has rejected the commonly accepted black-and-white view that trustworthy people are always trustworthy and untrustworthy people are always untrustworthy. It has been shown that this view leads to the conclusion that studying trust is worthless. This paper has also discussed various determinants and properties of trust on the basis of the idea that trust is not simply a matter of intention. In particular, it has shown that regarding trust simply as a matter of intention makes trust studies virtually impossible and excludes many interesting questions from analysis. All these discussions suggest the importance of expressing trust as a probability.

References

Arai, K. (1995), "Kurikaeshi Shujin no Dilemma Game niokeru Communication to Settoku [Communication and Persuasion in Repeated Prisoner's Dilemma Games]", *Ikkyo Ronso* [*Hitotsubashi Review*] 114, pp.996-1006.

Arai, K. (2000), "Koyo Seido no nakano Shinrai [Trust in Labor Market Institutions]," Keizaigaku Kenkyu [*Hitotsubashi University Research Series; Economics*] 42, pp.105-155..

Arai, K. (2005), "Game Jikken ni Arawareru Shiri Tsuikyu to Bunka [Pursuit of Self-interest and Culture Revealed by Game Experiments]," Keizaigaku Kenkyu [*Hitotsubashi University Research Series; Economics*] 47, pp.181-200.

Arrow, K. J. and G. Debreu. (1954), "Existence of an Equilibrium for a Competitive Economy," *Econometrica* 22, pp.265-290.

Barber, B. (1983), *Logic and Limits of Trust*, New Brunswick, N.J, Rutgers University Press.

Barney, J. B. and M. H. Hansen (1994), "Trustworthiness as a Source of Competitive

Advantage," *Strategic Management Journal* 15, pp.175-190.
Bohnet, I. and R. Zeckhauser. (2004), "Trust, Risk, and Betrayal," *Journal of Economic Behavior and Organization* 55, pp.467-484.
Bohnet, I., F. Greig, B. Herrmann, and R. Zeckhauser. (2008), "Betrayal Aversion: Evidence from Brazil, China, Oman, Switzerland, Turkey, and the United States." *American Economic Review* 98, pp.294-310.
Bolton, G. E. and A. Ockenfels. (2000), "A Theory of Equity, Reciprocity, and Competition," *American Economic Review* 90, pp.166-193.
Chiba, T. (1997), "Shijo to Shinrai : Kigyokan Torihiki wo Chushin ni [Markets and Trust: With Special Focus on Inter-firm Transactions]", Shakaigaku Hyoron [Japanese Sociological Review] 48, pp317-333.
Dawes, R. M., J. McTavish, and H. Shaklee (1977), "Behavior, Communication, and Assumptions about Other People's Behavior in a Commons Dilemma Situation," *Journal of Personality and Social Psychology* 35, pp.1-11.
Debreu, G. (1959), *Theory of Value*, New York, John Wiley and Sons.
Deutsch, M. (1962), "Cooperation and Trust: Some Theoretical Notes," in: M. R. Jones ed. *Nebraska Symposium on Motivation*, Lincoln, Nebraska, Nebraska University Press.
Engelmann, D. and Strobel, Martin (2004), "Inequality Aversion, Efficiency, and Maximin Preferences in Simple Distribution Experiments," *American Economic Review* 94, pp.857-869.
Fehr, E. (2009), "On the Economics and Biology of Trust," *Journal of European Economic Association* 7, pp.235-266.
Fehr, E. and K. M. Schmidt (1999), "A Theory of Fairness, Competition, and Cooperation," *Quarterly Journal of Economics* 114, pp.817-968.
Hong, K. and I. Bohnet (2007), "Status and Distrust: The Relevance of Inequality and Betrayal Aversion," *Journal of Economic Psychology* 28, pp.197-213.
Lazaric, N. and E. Lorenz (1998), "The Learning Dynamics of Trust, Reputation and Confidence," in: N. Lazaric and E. Lorenz eds., *Trust and Economic Learning*, Cheltenham, U.K., Edward Elgar, pp.1-20.
Nooteboom, B. (2002), *Trust: Forms, Foundations, Functions, Failures and Figures*, Cheltenham, U.K., Edward Elgar.
Ring, P. S. and A. H. Van de Ven (1992), "Structuring Cooperative Relationships between Organizations," *Strategic Management Journal* 13, pp.483-498.
Sako, M. (1992), *Price, Quality, and Trust: Inter-firm Relations in Britain and Japan*, Cambridge, Cambridge University Press.
van de Kragt, A. J., J. M. Orbell, and R. M. Dawes (1983), "The Minimal Contributing Set as a Solution to Public Goods Problems," *American Political Science Review* 77, pp.112-22.
Yamagishi, T. (1998), *Shinrai no Kozo* [The Structure of Trust], Tokyo, Tokyo Daigaku Shuppankai.
Zand, D. E. (1972), "Trust and Managerial Problem Solving," *Administrative Science Quarterly* 17, pp.229-239.
Zaheer, A., B. McEvily, and V. Perrone (1998), "Does Trust Matter? Exploring the Effects of Interorganizational and Interpersonal Trust on Performance," *Organization Science* 9, pp.141-159.