IESE Business School
University of Navarra

Working Paper

WP No 567

September 2004

BOUNDED RATIONALITY, VALUE SYSTEMS AND TIME-INCONSISTENCY OF PREFERENCES AS RATIONAL FOUNDATIONS FOR THE CONCEPT OF TRUST

Josep M. Rosanas *

* Professor of Accounting and Control, IESE

IESE Business School - Universidad de Navarra
Avda. Pearson, 21 - 08034 Barcelona. Tel.: (+34) 93 253 42 00 Fax: (+34) 93 253 43 43
Camino del Cerro del Águila, 3 (Ctra. de Castilla, km. 5,180) - 28023 Madrid. Tel.: (+34) 91 357 08 09 Fax: (+34) 91 357 29 13

BOUNDED RATIONALITY, VALUE SYSTEMS AND TIME-INCONSISTENCY OF PREFERENCES AS RATIONAL FOUNDATIONS FOR THE CONCEPT OF TRUST

Abstract

This paper intends to contribute to the (bounded rationality) foundations of trust. After reviewing the extant definitions, I establish the formal structure of situations involving trust. In that context, I examine the paradoxical situation of (calculative) trust in simple settings. Then I show how bounded rationality provides a rationale for a concept of trust that goes beyond that calculative notion. Value systems and possible inconsistency of time preferences are shown to be crucial elements.

Keywords: Trust, Bounded rationality, Value systems, Behavioral Decision-making.

BOUNDED RATIONALITY, VALUE SYSTEMS AND TIME-INCONSISTENCY OF PREFERENCES AS RATIONAL FOUNDATIONS FOR THE CONCEPT OF TRUST

Trust has been the focus of attention of many writers in recent years, from very different disciplines: economics (e.g., Arrow, 1974; Hirschman, 1984; Dasgupta, 1988; Kreps, 1990; Williamson, 1993; Casadesus-Masanell, 2004), sociology (Granovetter, 1985; Zucker, 1986; Coleman, 1990; Giddens, 1990), political science (Fukuyama, 1995; Misztal, 1998; Seligman, 1997), and management (Barney and Hansen, 1997; Shapiro et al., 1992; Mayer et al., 1995; Kramer and Tyler, 1996). Yet, as many scholars have noted, no common ground for the treatment of trust exists, in spite of various attempts in that direction. The importance of the subject, and the growing interest in it, are nevertheless widely recognized. Arrow (1974) was one of the first writers in economics to underscore the importance of trust in economic life:

> "Consider what is thought of as of higher or more elusive value than pollution or roads: trust among people. Now, trust has a very high pragmatic value, if nothing else. Trust is an important lubricant of a social system. It is extremely efficient; it saves a lot of trouble to have a reliance on other people's word. Unfortunately, this is not a commodity that can be bought very easily. If you have to buy it, you already have some doubts about what you've bought. Trust and similar values, like loyalty, or truth-telling, are examples of what an economist would call externalities" (p. 23).

Many analyses of trust are based on the underlying assumption that trust is rationally based and instrumental (Kramer and Tyler, 1996, p. 10). But this instrumental view of trust is not enough to explain its presence in social bodies. In fact, and again according to Kramer and Tyler, "trust is important only when people have social relationships" (1996:10). Besides, people often adopt some type of rule-based decision making, based on their own identities as individuals (March, 1999) and their identification with a group, possibly to protect existing social values and relationships, even in purely economic terms (Kahneman, Knetsch and Thaler, 1986).

Three recent entries in the field show the different approaches to the problem of trust. Korczynski (2000) attempts to integrate the sociological and economic approaches; James (2001) tries to show the contradiction of some possible interpretations of the rationalistic approach that is typical of the economics-based literature; and Casadesus-Masanell (2004) shows how ethical standards and altruism can arise in an agency theory context.

This paper intends to contribute to an integration of the different concepts of trust through the methods of formal analysis. It attempts to establish the decision-theoretical bases on which trust can be founded, showing that it follows from a rigorous definition and analysis of the problem that bounded rationality and value systems are essential for a concept of trust that goes beyond mere calculativeness.

The paper is organized as follows. First, I briefly review the definitions of trust that can be found in the literature, and establish the basic logical structure of situations potentially involving trust. Then, I proceed to examine simple settings where the notion of (calculative) trust is paradoxical: in fact, either there is no problem, or else there is no solution. Next, I show that uncertainty about preferences is an essential element in calculative trust, while the analysis of bounded rationality provides a reason for a concept of trust that goes well beyond the calculative notion discussed before. Value systems provide a basis for trust in integrity, and intertemporal consistency of preferences provides a basis for trust in character. Finally, I relate the instrumental-rational approach to trust to the social and cultural approaches.

DEFINING TRUST AND SETTING THE PROBLEM

Arrow, in the quotation that opened this paper, considered trust rather 'elusive'. Both Gambetta (1988) and Williamson (1993) used the same word ('elusive'), which implies difficulty in finding a satisfactory definition. Some authors make trust implicitly equivalent to truth-telling, or to keeping one's promises (Dasgupta, 1988), although Dasgupta attributes a slightly different meaning to the concept as well. Williamson (1993), on the other hand, does not consider trust to be an important concept: according to him, it is just another name for risk, and therefore all matters related to trust are mere calculation. Interestingly, though, in the same paper he accepts that, under the heading of 'personal trust', there is something involved in the word 'trust' that goes beyond mere calculation.

In a much more behaviorally oriented paper, Gabarro acknowledged that:

> "Trust has been defined or operationalized in the literature in many different ways, including the level of openness that exists between two people, the degree to which one person feels assured that another will not take malevolent or arbitrary actions, and the extent to which one person can expect predictability in the other's behavior in terms of what is 'normally' expected of a person acting in good faith" (1978: 294).

Thus, trust has to do with two economic agents. Zand, following Deutsch (1962), defined trust as "actions that (a) increase one's vulnerability, (b) to another whose behavior is not under one's control, (c) in a situation in which the penalty (disutility) one suffers if the other abuses that vulnerability is greater than the benefit (utility) one gains if the other does not abuse this vulnerability" (1972: 230).

Dasgupta stresses the second point: having the correct expectations about the other's behavior *before* that behavior can be monitored (1988: 51). Therefore, one agent makes the decision to trust first, and the other agent makes the decision of whether to honor that trust later on, with no monitoring.

Many authors have adopted the essence of Zand's definition – except perhaps point (c), which is not crucial to many aspects of the analysis. Kreps, for instance, implicitly adopts a similar concept when he presents the analysis of trust as a "one-sided version of the prisoner's dilemma game" (1990: 101). Mayer, Davis and Schoorman, in a paper that is intended to be integrative of previous research, define trust as:

> "the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party" (1995: 712).

Barney and Hansen adopt a similar point of view: "an exchange party worthy of trust is one that will not exploit the other's exchange vulnerabilities" (1997: 6), which is practically the opposite of Williamson's notion of *opportunism*, as the willingness to profit at the expense of others, or "a condition of self-interest seeking with guile" (1985, p. 30). Korczynski recognizes that in "much of the growing literature on trust the concept is intimately tied in with vulnerability or risk"; and that the literature "is replete with categorizations of types of trust, from rational calculative trust, altruistic or blind trust, to distinctions made between personal trust, and trust in abstract systems and institutions..." (2000:3).

Basic trust structure

I will adopt the Kreps formulation, which is consistent with the essential points that the above definitions have in common. The structure of that formulation is based on the following points:

1) The situation involves two decision makers, A (the trustor) and B (the trustee). The decision-making process is sequential: A makes the first decision (whether or not to 'trust' B); and, if the decision is 'trust', B can make the second decision ('honor the trust' or 'betray'). This second decision determines the monetary consequences affecting the two decision-makers. If A 'does not trust', B can do nothing and they both get zero.

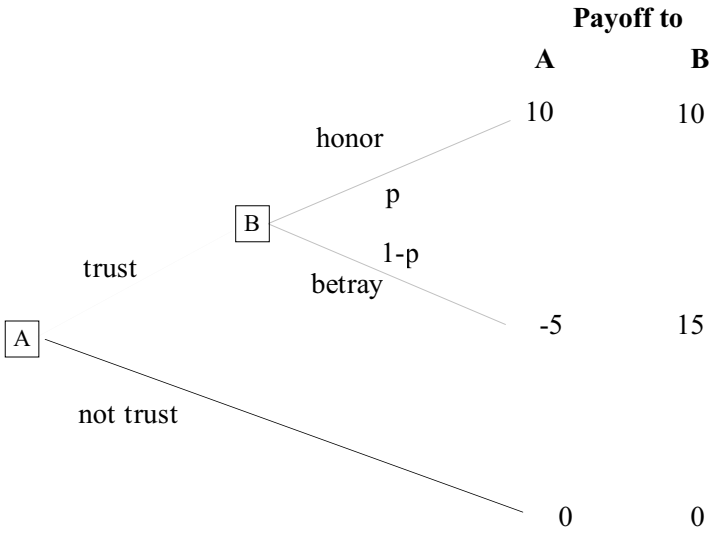
2) The definitional idea that A 'makes him/herself vulnerable to B' is formalized as follows: If A takes the alternative 'trust' and B 'betrays', then A's payoff is negative; while if B 'honors', A's payoff is positive. Therefore, A can gain by trusting B if that trust is honored, but makes him/herself vulnerable to a loss if it is not.

3) B's payoffs, for the problem to be interesting, must go the other way around: they should be greater if B 'betrays' than if B 'honors' (though positive under both alternatives). If this was not the case, the interests of A and B would be perfectly aligned, there would be no possible conflict and, therefore, no need for trust.

4) In general, for A, there is uncertainty as to whether B will make the decision to 'honor' or to 'betray'. The (subjective) probability (to A) that B will 'honor' is denoted by p. Of course, the probability of B 'betraying' is $1-p$. As we will immediately see, under the preceding assumptions, and if the game is played only once, p can only be zero. But in other situations, it might not. Explaining where this p comes from in general situations is one of the major objectives of this paper.

This formulation is shown as a decision tree in Figure 1, where the first node (the square with an 'A') represents A's decision in the usual sense, while the second node (the square with a 'B') represents B's decision. Thus, from A's point of view, it is an uncertainty node, to which A can assign a probability distribution.

Figure 1



Initial assumptions

I will assume throughout that the monetary results are 'fixed' and cannot be changed by the interaction of the two agents or by any superior authority. This excludes incentive systems designed by superior authorities, penalties imposed by such authorities to enforce possible agreements, or side-payments between the agents. The reason for this is simple: if payoffs can be changed by someone in order to achieve better results, those better results will not be based on 'trust' but on the new payoffs, and therefore on 'authority'. Or, in other words, conditions 1) to 3) above will not hold with respect to the actual payoffs established by such a system.

Besides, I will assume that before deciding on what action to take, agents can freely communicate with each other and make their decisions on the basis of any conceivable agreement (except, of course, those involving transfers of money or monetary equivalents between them, prohibited in the above paragraph). As I will show, under bounded rationality, 'persuasion' may be an important consequence of this communication process.

Other assumptions will change as I develop my argument. To be specific, there are four possible assumptions that may substantially affect the results of the analysis:

1) The two agents may value exclusively money or monetary equivalents, or they may value also some other, perhaps intangible variables, such as a job well-done, reputation, learning, other people's welfare, keeping one's word, and so on.

2) Agents may be certain or uncertain about the payoffs that will obtain under each alternative. Figure 1 assumes certainty, but it can obviously be generalized to the case of uncertainty, where payoffs depend on a state variable.

3) Agents may have bounded or unbounded rationality. With unbounded rationality they have perfect preference orderings, and are able to foresee with precision the probability of any uncertain circumstance, while with bounded rationality they may not be sure about which alternative is best for them, or may have preferences that are not entirely consistent.

4) A trust situation like the one in Figure 1 may take place only once, or it may be repeated (not necessarily with the same exact payoffs, or in the same circumstances) several, or even an infinite number of times.

WHERE IS TRUST UNDER UNBOUNDED RATIONALITY?

The simplest scenario

The simplest possible situation is one where the payoffs are known to both agents with certainty, agents value money exclusively, have unbounded rationality, and are certain that the situation will occur only once (i.e., it will not be repeated).

Analyzing that tree is straightforward. If the only thing that matters to the agents is money, B, if 'trusted', will choose 'betray'. A, knowing this, will assign zero value to the probability of B choosing 'honor' and, therefore, will choose not to trust B. Both agents will then end up with zero utility. Of course, this is a Pareto-inferior solution to 'A trusts & B honors', where both would get positive monetary rewards. But playing the game one time only, and having utility for money only, that is the only feasible alternative.

Notice that for the problem to be meaningful, the payoff to B in the event that A 'trusts' and B 'honors' has to be positive. If it is negative, no matter how small, then the final outcome would still be the status quo (A does not trust & B does nothing), although that outcome would now not be Pareto-inferior to the solution 'A trusts & B honors'. The status quo could then be considered as a 'compromise' between the two other possibilities, one favoring the first agent and the other, the second.

Of course, if the payoffs to B are reversed (i.e., B gets better results by 'honoring' than by 'betraying'), then there is no problem: it is in B's interest to 'honor' A's trust. But then we can hardly speak of trust as a concept, but rather of an incentive system that yields the right results to both agents.

Proposition 1*: In the simplest scenario (utility only for money, unbounded rationality, certainty about the outcome of each other's actions, and playing only once), if an agent A can choose between 'trusting' the action of another agent B, then, depending on B's payoffs, either there is no place for trust (because the incentives are perfectly aligned) or else trust cannot exist: both agents will end up in a situation that is Pareto-inferior to the one that is theoretically feasible.*

Or, in simpler terms, either there is no problem, or else there is no solution. The reasons for trust to exist, then, will have to be found in different assumptions.

The trust paradox

The solutions typically found in the literature to achieve Pareto optimality change, as I will show, at least one of the conditions of the simplest scenario above. James (2002) provides an insightful summary of those solutions: (i) writing explicit contracts; (ii) repeating the interactions (repeated game); (iii) relying on implicit contracts; and (iv) changing the preferences.

The first two solutions, as has been suggested already, essentially consist of changing the payoffs so that both agents have an incentive to choose the alternatives that lead to the Pareto-optimal solution. Writing explicit contracts is in essence a way to change (voluntarily and ex-ante, of course) the monetary payoffs to the individuals so that they have an immediate incentive, compatible with the other party's incentive, to act in a way that is Pareto-optimal. Thus, the original problem is replaced (through an agreement between the agents) by a different one, in which both agents have the right incentive.

The repeated game solution can also be seen as a way to change the monetary payoffs to the individuals. In a situation that is going to be repeated a number of times, a tit-for-tat strategy followed by both players changes the payoffs to the agents. Specifically to B, whose choice is crucial: the payoff if B betrays remains the same (15), but the payoff if B honors is replaced by the present value of the former payoff (10) an indefinite (perhaps infinite) number of times. Thus, B has an incentive to 'honor', and A, knowing that, will 'trust'. Kreps shows how, besides, certainty about having future interactions is not necessary for that result: provided that the probability of a new interaction is 'big enough', it is in the best interest of B to 'honor', and therefore it is in the best interest of A to 'trust' (1990: 102).

Therefore, when the only variable under consideration by the agents is the monetary payoff, in both of the first two solutions to the problem of trust the concept simply vanishes because it is made unnecessary: if both parties have the incentive that leads to the Pareto-optimal solution, then there is no need for trust. This is what James (2002) calls the 'trust paradox'.

The third and fourth solutions (establishing implicit contracts, and changing the individuals' preferences), can be seen as changing the payoffs as well, but changing the subjective value of the payoffs instead of the objective monetary payoffs. In both cases, the possible solution necessarily implies that the agents value variables other than the economic ones; which goes beyond the scenario considered so far. As I will show, trust acquires full meaning only when other, non-measurable or non-economic variables are taken into account; discussion of those other variables is therefore deferred until we examine that case below.

Thus, using the results of the first and second solutions we can state the following proposition:

Proposition 2*: If agents value only monetary payoffs, explicit side contracts and repeated-game solutions are just a way of modifying the payoffs so that the need for trust is eliminated. Therefore, they do not really constitute a solution to the trust problem, but rather a solution to the optimality problem.*

Incorporating uncertainty: trust in judgment and competence

The minimum requirement to add to the previous scenario to provide some foundation to the notion of trust is uncertainty. To avoid getting into issues relating to risk and insurance, which have been extensively analyzed in the principal-agent literature, I assume that both A and B are risk-neutral. So, if both A and B agree on the probabilities of the different states of nature, the nature of the problem is exactly the same as in the simplest scenario, but with the expected values of the actions instead of the certain values we had before. Therefore, again, either there is no problem, or else there is no satisfactory solution.

The situation changes completely if A and B do not agree, i.e., information about the states of nature is asymmetrical. Then, even if the payoffs to both agents are identical, and so there is no conflict in that respect, some notion of 'trust' is needed for A to delegate the decision to B, because given the probability distributions, the possible results may look different to the two agents: A may prefer one alternative while B prefers the other (see **Appendix 1** for a simple numerical example).

The reasons why A might consider trusting B typically have to do with decentralization and specialization. It may be necessary for A to delegate the decision to B, for instance, because the decision is made and implemented at the same time, and A cannot do it him/herself, or because of A's information overload. Then, if A believes that B is an expert in this kind of decisions and therefore 'knows better' than him/herself, A may 'trust' B's judgment and accept that he/she makes the decision, which makes A vulnerable to a loss of 5. The alternative is probably 'doing nothing', and therefore risking zero, but with zero return. Aghion and Tirole (1997) have shown how the transfer of authority from the principal to an agent increases the agent's initiative to acquire information, thus reducing the principal's information overload, which makes 'trusting B' an even better alternative.

The reasons for trusting B's information (i.e., his/her probability assessments) are essentially empirical. Since, in our setting (sharing the results), there is no possible conflict of interest between the two parties, both A and B want 'the right decision' to be made. And A will trust B if B has a record of good decisions that show him/her to be an expert in making this type of decisions.

Some of this empirical evidence is often summarized in the standards accepted by a given profession; and, therefore, A might we willing to trust B, without direct empirical evidence, if B adheres to the standards of that profession: accreditation, commonly accepted practices, and so on are the usual ways. But still, the basis of that trust is essentially empirical. The basic ideas of the last two paragraphs, then, lead us to the following proposition:

Proposition 3*: Uncertainty and asymmetrical information provide a sufficient condition for the concept of trust to be meaningful. If one agent believes that another agent has better information than him/herself, the first agent may 'trust' the second agent's assessment in order to make the decision.*

Non-measurable variables: trust in preferences

The situation becomes more complex if both A and B have utilities for other intangible and non-measurable variables, like 'effort aversion', 'effort love', 'fun when doing something', 'prestige', 'recognition', 'loyalty', 'friendship', or whatever. Typically, the set of all these variables and their impact on the (perceived) well-being of the decision-maker include what has been called 'intrinsic motivation' and possibly other, altruistic motives. We now continue to assume that both agents have unbounded rationality, so that they know their own utility functions perfectly and are able to evaluate the possible alternatives without doubt, and without possibility of 'changing their minds'. Also, for the time being, the assumption will be that such preferences are 'given'.

Assume that B's utility function is perfectly known to A (notice that it is in general irrelevant whether the converse is true, because A plays first). Then, the problem is formally identical to the one studied in the simplest scenario analyzed before. If the total utilities to A

and B can be represented by the values in Figure 1 (including both economic and non-economic variables), again there is no satisfactory solution nor place for trust. It will not matter whether the monetary results are aligned with total utility or not: total utility is all that matters in a world of unbounded rationality and perfect information. And, as in the simplest scenario, depending on B's total utilities, either there is no problem or else there is no satisfactory solution.

There is a particularly interesting case, though, in which the agents' utility for other, qualitative variables may make possible a 'solution' to a situation which in terms of the monetary variables alone is precisely our simplest scenario. If that is the case, in terms of these variables alone, we have just seen that a Pareto-optimal solution is impossible. Now, if the results in terms of total utility are such that the utility to B of honoring is in fact greater than the utility to B of betraying, then the 'A trusts & B honors' alternative is Pareto-optimal both from the point of view of the economic variables alone and from the point of view of total utility.

In such a case, non-economic motives may make it possible to reach an efficient solution that would be impossible with economic variables alone. A numerical illustration of this possibility is shown in **Appendix 2.** With unbounded rationality and perfect knowledge of the utility values, the choice is clear to both agents. Therefore, we are back to the 'trust paradox', where by including non-monetary variables and preferences perfectly aligned, we have made real trust unnecessary. Or, perhaps, using the words in a slightly different way, one might say that in this type of situation, A 'trusts' B's preferences 100%, i.e., A can be sure that B will 'honor' and that thus they can avoid an economically inefficient solution.

The previous analysis leads to the following proposition:

Proposition 4*: When agents value non-economic or non-quantitative variables, but have unbounded rationality and perfect knowledge of the utilities, then the analysis is quite similar to that of the simplest scenario: trust is either unnecessary (if utilities are aligned) or else cannot be obtained.*

A special case: trust in unselfish values

A particular case of preferences about intangible, non-measurable variables is when agents care for each other's welfare. Formally, we can express this as each agent's utility being an argument of the other agent's utility function. A numerical illustration of this possibility is shown as **Appendix 3**.

There, it has again been assumed that both agents have full knowledge of the other agent's utility. Hence, the probability that B will 'honor' A's trust is either 0 or 1. Therefore, the word 'trust' applied to this situation is just as paradoxical as before: knowing that the other agent will do what is best for him/herself hardly coveys the intuitive idea of 'trust'. It might be argued, though, that there is a qualitative difference between the two types of reasons for 'trust' involved in the two previous sections: 'trusting' individuals because of their preferences with respect to whatever non-measurable variables (say, the type of work they like, reputation, or environmental factors) is less related to the intuitive idea of 'trusting' than 'trusting' individuals because their preferences include our own well-being. The latter case, at least, includes an attitude of benevolence towards another individual that we usually associate with trust. But, as I will try to show next, trust becomes partly meaningful only

when there is uncertainty about the other person's preferences, and acquires full meaning only in the context of bounded rationality.

It is interesting at this point to contrast these results with some notions found in the literature. For instance, Shapiro, Sheppard and Cheraskin (1992), distinguish between three kinds of trust: deterrence-based, knowledge-based, and identification-based. Both deterrence-based and knowledge-based trust fall entirely into the trust paradox. In the organizational context, they suggest that deterrence-based trust can be achieved by three means: repeated interactions, multiple interactions, or taking hostages. In any case, the result is a change in the actual payoffs to individuals, so that it is in their best interest to act in the interest of the organization. Knowledge-based trust, in these authors' interpretation, consists of knowing sufficiently the other party's preferences so that an incentive (possibly monetary) can be established that is enough to persuade the other party to behave as desired. Identification-based trust, in contrast, would then be exactly as analyzed in this section, except for the fact that the authors suggest several ways to achieve that identification (name, shared values, proximity, etc.) that can exist only in the context of bounded rationality; and therefore we defer discussing them until we explicitly consider that concept.

Portales, Ricart and Rosanas (1997), along similar lines, include deterrence-based and knowledge-based trust in 'calculative trust', which again falls into the trust paradox. Beyond calculative trust, they analyze integrity-based trust and personal trust. Integrity-based trust is based on the trustee's preferences and values, while personal trust is based on altruistic motives and an imperfect knowledge of one's own preferences. Again, as we will show, this can only be rigorously analyzed in the context of bounded rationality. The second part of this paper can be considered an expansion of their analysis, explicitly introducing bounded rationality.

Imperfect knowledge of other agents' preferences

Assume now that B's utility function is *not* perfectly known to A. Then, the (however imperfect) knowledge of A about B's preferences will be reflected in A's subjective probability p of B honoring A's trust. How likely is it that other dimensions of the results of the problem (possibly including A's monetary rewards) are more (or less) valuable for B than hard money? It is this imperfect knowledge that makes this probability, in general, to be between 0 and 1.

With unbounded rationality, however, A is able to assess probability p exactly and without ambiguities. A's rational answer to the problem, therefore, is one of mere calculation. Given the probability p of B honoring A's trust, is A better off trusting B, or not trusting him/her? Williamson's (1993) notion of 'calculative trust', based on the seminal paper by Deutsch, in which he claims that trust is just another name for risk, clearly applies to this type of situation.

It is interesting to reconsider here the James (2002) distinction, following the analysis by Lahno (1995), between 'trust as prudence' and 'trust as hope'. In this paper's framework, if the assessment by A of the probability p of B honoring the trust is exact, with no fuzziness or ambiguities, and if the computation is favorable to 'trusting' B in monetary terms, A can be said to trust out of 'prudence', which would be equivalent to Williamson's 'calculativeness'. If, with the same assessment, the computation is unfavorable to A in monetary terms, A can decide to 'trust' B out of a (non-rational) 'hope' that *this time* B will

honor his/her trust. Of course, this hope is 'irrational', according to Lahno and James; or, perhaps, one might say it is just as rational as the compulsive gambler's belief that *this time* he/she will win. 'Trust as hope', then, requires some bounds to rationality at least.

Alternatively, A can make his/her decision based on the belief that, in the future, it will help create an atmosphere of trust in the environment such that the different A's and B's will trust each other more; and that, in turn, would have to do with future interactions. But this goes well beyond the mere 'tastes' usually represented by a utility function, and possibly includes social and cultural elements beyond the rationalistic view.

The knowledge necessary to assess p, obviously, comes from previous contacts (direct or indirect) between the two individuals involved, perhaps in many other previous situations completely different from the 'trust problem' analyzed here. I will go back to the determination of p in more detail later on; suffice it now to say that in absence of any specific information, probability p would be the 'a priori' probability that the average person with whom A interacts can be trusted. Then, in a world of unbounded rationality, that person would be able to assess the probabilities of the signals, etc., and proceed to a Bayesian updating of probability p.

REVISITING BOUNDED RATIONALITY

Ill-known payoffs and preferences

Most analyses of bounded rationality focus on the limited ability of human beings to derive (or calculate) the consequences of their actions. Herbert Simon's original formulation of the concept, though, included three characteristics of human thought. The first is of course the limited ability of human beings to foresee the consequences of their actions and the logical implications of their thoughts. But also, human beings have limited ability to anticipate how they will like the consequences of their actions, or how much they will like the action itself: "It is a commonplace experience that an anticipated pleasure may be a very different sort of thing from a realized pleasure" (Simon, 1997, p. 95). Finally, they have limited ability to find possible courses of action as solutions to problems.

To analyze the role bounded rationality plays in the problem of trust, we basically need the first two of these limitations; and it is in fact the second one that has greater implications for that purpose. Let us proceed to analyze them in turn.

The first one essentially means that, in complex situations, agents may not be able to know the actual payoffs to themselves, not only because of 'external', or 'objective', uncertainty (which may be present, of course), but because they are not able to figure them out, or determine their probabilities with any degree of accuracy. Being able to accurately assess the probabilities of uncertain external events is one of the characteristics of unbounded rationality. The inability to assess such probabilities includes the possibility of being unable to foresee certain circumstances; or, in more formal terms, the possibility that agents may assign a zero probability to events that are perfectly conceivable, because they have simply 'overlooked' them. For A, this limitation represents uncertainties with respect to B's behavior in addition to the usual uncertainties about 'objective' events.

The meaning of the second characteristic is like an extension of the first one: agents are unable to predict accurately the *subjective* part of the payoffs, i.e., the utility to themselves of the expected results. In other words, decision-makers have to anticipate future preferences of which they cannot be sure. March states this distinction very clearly:

> "The conception of choice enshrined in the axioms of contemporary decision theory and microeconomics assumes optimization over alternatives on the basis of two guesses. The first guess is about the uncertain future consequences that will follow from alternative actions that might be taken. The second guess is about the uncertain future preferences the decision-maker will have with respect to those consequences when they are realized" (1987:155).

In the context of trust, the implications go beyond the uncertainty added to A about B's behavior that has already been considered above. The problem is particularly interesting when the explicit, quantifiable payoffs are of the kind that lead to the trust problem, i.e., like the monetary payoffs in **Appendix 2**. B, then, has an explicit incentive to 'betray'. But what is the value to B of the non-quantitative variables? Under unbounded rationality, the answer is quite clear (at least to B). Thus, if the non-quantitative variables are valued as in **Appendix 2**, the problem ends there, as we argued when analyzing this case. In contrast, with bounded rationality, it may well be that honoring A's trust is in the best interest of B, but B him/herself may not know it (or, at least, not without some doubts or fuzziness).

Also, under bounded rationality, B may be sorry *after* making his/her choice, whatever that choice may be, which may change the alternative B chooses next time, if there is a next time. For instance, B may at the moment of making the first decision 'magnify' the importance of the quantitative variables and 'betray' A, only to realize –too late– that the non-quantitative variables were at least as important as the quantitative ones. Of course, the opposite may happen as well: on reflection, B may decide to 'honor' A's trust, only to find later on that the non-quantitative variables were not that important after all.

Human beings may face the same problem over and over again through time and make different decisions because of learning, or because of different states of mind (emotion, for instance) that bring into the focus of their attention some aspects of their lives and some of their values to the relative neglect of others (Simon, 1983; Simon, 1987; Loewenstein, 1996). In the presence of a time constraint (which would be irrelevant in the case of unbounded rationality) when decision time is scarce, an optimization approach may not be feasible for an agent who is not-so-familiar with the problem (Selten, 1999).

The ability to optimize may not be symmetrically distributed in a problem involving trust (i.e., one of the agents may be more familiar with the problem and how to solve it than the other). To be specific, if A is more familiar with the problem than B, A may think it a possibility that B will choose 'wrongly' according to B's own preferences, harming A in turn. The possibility of A trusting B is obviously affected by this type of assessments.

Hirschman (1984) has argued that there are two kinds of activities. Some human activities are instrumental, and are done in order to get a paycheck or an explicit reward. But others are not: those activities that are undertaken 'for their own sake' or that 'carry their own reward' fall into this category. Some activities have such an uncertain reward that they will seldom be undertaken because of that. But there is an 'education process' in such activities: not everybody likes them, only those people that have 'learned to love them'. This is unlikely to happen in simple, routine types of jobs or activities. But it is much more likely in more complex situations, precisely where trust is of higher importance than in rather simple

contexts. The Hirschman analysis includes other factors, like the willingness to put plans of action into practice, which will be considered below.

In general, preferences may change through time, depending on each agent's experience. An agent may 'learn to love' some variables or situations, and 'learn to hate' others. Intangible variables found in business situations, like the value placed in personnel development, the public image of the company, or the internal human climate of the firm, may change dramatically through time, as agents learn about their jobs, about the organization and about themselves.

An important aspect of this learning process is persuasion. In a world of unbounded rationality, there is no place for persuasion: every agent knows perfectly his/her preferences (including the evolution of preferences through time), and there is no reason to change. In contrast, in a world of bounded rationality, while some variables (monetary rewards, for instance) are easy for everybody to appreciate, persuasion may play an important role in 'learning to love' some goals or activities. Barnard (1938) already stressed the importance of persuasion as one of the crucial methods (together with incentives) to get people to work in the interest of the organization. Nowadays, common everyday experience, as well the momentum of communication courses in business schools, can be said to confirm Barnard's intuition.

The importance, for the analysis of trust, of the imperfect knowledge of one's own preferences, which implies the kind of learning just mentioned, cannot be overstated. In a situation like the one shown in **Appendix 2**, the monetary rewards are perfectly known, and their utility to the agents obvious; but the total utilities on the right-hand side of the Table may depend (under bounded rationality) on whether someone has used some persuasion on the other party. B, for instance, may be rather indifferent about any one (or several) of those variables; but A may try to persuade B that such variables are worthwhile. If A believes B has been persuaded, A may 'trust' B to achieve a result that is optimal economically and otherwise. But, of course, if B is disappointed, his/her attitude next time around will surely be different. So persuasion needs some element of truth to work in repeated interactions.

Systems of preferences and values

Bounded rationality is, according to Simon, the type of behavior that is intendedly rational, but only limitedly so. Typically, it is 'intendedly rational' through a system of preferences and values. Preferences of individuals are considered in economic theory to be 'arbitrary'. Economic theory assumes rationality from the point of view of the consistency of those preferences to avoid circularities; but except for that aspect, the preferences of two individuals between, say, two goods, may be completely unrelated. In a world of unbounded rationality, a preference map is a perfect reflection of the individual's preferences and values; and there is no distinction between different levels of values by importance, or by familiarity with the specific situation.

Under bounded rationality, however, values and preferences may look substantially different. Herbert Simon borrowed from the logical positivistic philosophy (Simon, 1997 ed., chapters 1, 3 and 4) the distinction between fact and value; and, in accordance with that philosophy, he related rationality to the choice of means conducive to the achievement of previously selected goals (1997, p. 4). The selection of goals itself would then not be rational or irrational, it would just be a matter of taste on the part of the individuals.

Simon recognized, though, that there is a 'hierarchy of decisions', a means-ends chain so that a given goal is often a means to achieve a higher end: "ends themselves are often merely instrumental to more final objectives". "Rationality, then, has to do with the construction of means-ends chains of this kind" (Chapter 4, p. 73). Therefore, whether a final end is achieved through an intermediate end or not is entirely a matter of fact, susceptible of being empirically tested. For many human beings, most of the means pursued actively are only means to higher ends.

Preferences, then, operate essentially on the higher ends; and the lower ends are obtained as a consequence. Rokeach (1973) claims that people have relatively few basic values. Then, according to Fischhoff, Slovic and Liechtenstein,

"If, as Rokeach claims, people have relatively few basic values, producing an answer to a specific value question is largely an exercise in inference. We must decide which of our values are relevant to the situation, how they are to be interpreted, and what weight each is to be given." (1988: 401)

This applies in particular to the context of the results to be achieved in a business firm, or in any organization in general. Profitability (or financial equilibrium) is always one of the crucial variables desired by managers, but at the same time they also value 'market position', or 'competitive advantage', or 'organizational knowledge', perhaps as a means to achieve the higher end of profitability, or perhaps even as higher ends themselves. It is often claimed that, in business firms, profitability is the overriding goal and that all other ends have to be considered 'intermediate', and evaluated instrumentally as means.

But individuals may not be able, or willing, to make decisions in accordance with the relatively few basic values. Bounded rationality limits their ability to do so, and they have spontaneous impulses that may not go in the same direction as these few basic values. Hence, their ability to make choices that are logically consistent with them is also limited. However, these limitations differ depending on the situation: in familiar situations, individuals are better able to optimize with respect to higher ends, and, thus, they may have very definite preferences in relation to lower ends:

> "People are more likely to have clear preferences regarding issues that are familiar, simple, and directly experienced. Each of these properties is associated with opportunities for trial-and-error learning, particularly such learning as may be summarized in readily applicable rules or homilies. Those rules provide stereotypic, readily justifiable responses to future questions of values. When adopted by individuals, they may be seen as habits; when adopted by groups, they constitute traditions" (Fischhoff, Slovic & Liechtenstein, 1988: 399).

This may be seen as a specific instance of Simon's view about the role of intuition and emotion. Simon considers intuition as a 'shortcut' in the chain of reasoning, in situations that are familiar, or that 'ring a bell' on similar experiences from the past (Simon: 1987). The Fischhoff et al. quote above may thus be seen as the application of that approach to the selection of values: for familiar situations, individuals know (or think they know) what preferences are coherent with the higher values; but in more unfamiliar situations, they might not. Fischhoff et al. also provide an interesting list of the states of mind associated with not knowing what you want in less familiar situations and some actions that follow (1988: 400). That list is shown as **Appendix 4**.

We can see there that having a coherent opinion and accessing it properly (the implicit assumption in fully rational models of behavior) is only one possibility among many. Knowing what you want is, one might say, almost the anomaly. It is perfectly possible for someone to have no opinion, or have an incoherent one, not to realize it, and make decisions is spite of that. Living with incoherence is another possibility: in spite of higher ends being incompatible with some lower ends, people may try to achieve both (and fail, of course).

Applied to a situation involving trust between two people, the exercise in inference suggested by Fischhoff et al. in the above quotation would essentially consist in evaluating the possible dimensions of the consequences of each alternative, and evaluating to what extent the higher values are served. But the action to be initiated following this analysis may be in contradiction with lower ends, or immediate desires, or impulses, and the individual may be willing to be inconsistent. For instance, an individual may evaluate that Alternative 1 is better than Alternative 2 in terms of the higher values, but not in terms of the immediate monetary rewards: if those of Alternative 2 are bigger, the individual may be willing to live with that inconsistency.

In the next section we will go into the problem of willpower, by which one individual may really want to achieve a high objective, but be incapable of taking the corresponding action in the short run. For the time being, we are not analyzing that problem yet, but only the problem at the cognitive level.

It follows from the list in **Appendix 4** that rationality is not equally bounded for everybody. That is, some individuals try harder to be rational than others: some people are more 'reflective', and willing to check for coherence and act coherently with the higher ends, and others are more 'impulsive', or willing to pursue immediate, lower ends without much reflection or need for coherence. And for any specific individual, the probability that he/she is going to be more coherent is higher in familiar issues than in not-so-familiar issues, as stated in the above quotes by Selten and Fischoff et al.

In summary, some people are willing to act coherently with their stable, higher values in spite of their short-term urges to do otherwise, and some are less willing. The type of behavior that consists of trying to be coherent with some stable, or permanent, set of higher ends is what is usually called 'integrity'. According to Webster's Seventh New Collegiate Dictionary, integrity is precisely the 'adherence to a code of moral, artistic or other values' (Webster: 439). An important part of an individual's trust in other individuals will then reside in their integrity. To this type of trust we now turn our attention.

Trust based on integrity: value systems

If we reexamine the trust problem in the light of the previous section, we see that trust acquires a different meaning. The trustor, A, can make an assessment of the basic values of the trustee, B, and his/her willingness to make decisions consistent with that set of values. This second part, of course, is what makes this situation different from the case of unbounded rationality analyzed before: the first part is the same, as some knowledge of the other individual's set of values and preferences is necessary. With unbounded rationality, though, consistency between values, and the preferences expressed in any action, exists by definition; with bounded rationality, in contrast, there is an additional uncertainty about B's integrity, i.e., B's capacity to make decisions consistent with that basic set of values and preferences.

It is important to note that no assumptions have been made so far on the content of the basic values. Thus, those values may be shared by A, or not. A may trust that B will (or will not) do something because of the basic set of values A assumes B has, not because A agrees with those values. Suppose, for instance, that A and B have different religious beliefs. A may know that B's religion forbids some practices, and that B is a strong believer. Then, A may 'trust' B even if that action, forbidden by B's religion, is not considered 'bad' by A at all. The example of Yudhishthira in the *Mahabharata* epic, cited by Dasgupta (1988), is very much to the point here. Renowned for his trustworthiness, Yudhishthira lied once to throw off his unrighteous enemies. The lie worked because the enemies (having a completely different value system that did not value trustworthiness) believed Yudhishthira would not lie.

In the means-ends chains that human beings construct under bounded rationality, intermediate values are means to more final values. Obviously, though, those higher values need to be consistent with each other; otherwise a contradiction would require sacrificing one of them. And, in specific contexts, contradictions may arise. In this context, the Mahabharata example is particularly useful, because Yudhishthira would not lie to obtain immediate benefits for himself. Truthfulness is to him an important (higher) value; but he lies to defeat unrighteous enemies; and, thus, according to Dasgupta, qualifies as a consequentialist. A slightly different way of looking at the same problem is as a conflict between two values both of which are rather high in the hero's preferences: truthfulness and the good of his people, perceived as being incompatible at some point in time, and resolved in favor of defeating the enemies even if it is through a lie, because the good of his people is 'higher'.

The basic set of values may be more or less volatile, depending on specific individuals and on their circumstances. But A's beliefs about B's set of basic values (the content of those values, how 'stable' they are, and to what extent B is willing to make decisions based on them) will determine A's subjective probability p that B will honor or betray his/her trust. Then, 'trusting' B means in this context A's belief that B is willing to make a given decision according to his/her entire set of values, beyond the explicit, monetary values involved in the decision. That is the concept of trust based on integrity, which can be expressed in the following proposition:

Proposition 5*: Under bounded rationality, an essential element of trust in specific situations is the trustee's willingness to act in accordance with his/her system of values and preferences, because this makes the behavior of the trustee more predictable in such specific situations.*

No matter what the values are, and whether the trustor agrees with them or not, a person may be willing to make decisions consistently with them, or, on the contrary, may be rather impulsive. But, then, the result in terms of trust depends on the specific situation and the specific values involved. In other words, we cannot say that 'A trusts B' in general, but 'A trusts B' only under some specific circumstances and in some specific issue. Of course, the fact that the behavior of the trustee is predictable does not necessarily mean that the trustee will generally make decisions that are favorable to the trustor. For this to be true, we need to go one step further, to consider the kind of values that trustor and trustee have.

From this point of view, values can be 'purely selfish' or 'altruistic'; they may rate truth-telling, or fairness, or friendship, or social welfare, or the common good (the 'summum bonum' of the Schoolmen) high or low. If B's values are 'selfish', then A can perhaps trust B about a specific problem, or for a small class of decisions only. In contrast, if B's values are 'non-selfish', A may be willing to trust B for a wide class of decisions.

Proposition 6*: Under bounded rationality, an essential basis of trust between two people under a variety of circumstances is that the trustee must have a system of values and preferences in which some 'non-selfish', or 'social', values (or, simply, the interests of other people) are placed high, and be willing to make decisions according to such a system of values and preferences.*

The problem of self-command: can you trust yourself?

In economic models of decision-making and organization (with unbounded rationality), it is typically assumed that people are impatient, i.e., that they like to experience rewards soon and costs later. This is captured in such models through the use of a utility function discounting utility over time exponentially. Such preferences are called time-consistent. But, in O'Donoghue and Rabin's words,

> "Casual observation, introspection, and psychological research all suggest that the assumption of time-consistency is importantly wrong. It ignores the human tendency to grab immediate rewards and to avoid immediate costs in a way that our 'long-run selves' do not appreciate" (1999: 103).

The problem of the discrepancy between a person's preferences at different times is also an old one in philosophy. The basis of Aristotle's criticism of Socratic ethics was that very point. In Aristotle's analysis, Socrates had said that "nobody acts in opposition to what is best if he has a clear idea of what he is doing. He can only go wrong out of ignorance". The reasoning, however, and again according to Aristotle, "is in glaring contrast with notorious facts": people may know what is right and not do it because of weakness of will, or lack of control (Aristotle, Nicomachean Ethics, Book VII).

Schelling (1978; 1984) has studied the problem of time-inconsistency in depth. Wanting to quit smoking but not doing it; Christmas accounts that protect your money from yourself; free loans from the taxpayer to the IRS by understating the number of dependents; placing the alarm clock across the room; setting the watch a few minutes ahead to deceive oneself…

> "In these examples, everybody behaves like two people, one who wants clean lungs and long life and another one who adores tobacco, or one who wants a lean body and another who wants dessert. The two are in a continual contest for control: the 'straight' one often in command most of the time, but the wayward one needing only to get occasional control to spoil the other's best laid plan" (1978:290).

The 'two selves' are not equally important in Schelling's analysis. He is obviously partial to the 'straight' self. The section titled "Strategy and Tactic", for instance, in the 1984 paper, consists of recommendations so that the 'straight' self can be in command of the 'wayward' self: relinquish authority, let somebody else hold your car keys, order your lunch in advance, don't keep liquor, or tobacco, in the house, order a hotel room without television, do your food shopping after breakfast.....

Bazerman, Tenbrunsel & Wade-Benzoni (1998) formulate the problem in a slightly different way. They suggest that the two-selves theory can be conveniently made easier, into a 'want/should' explanation, based on the empirical evidence available. When people are asked what they want, their responses will be emotional, affective, impulsive, and hot-headed; whereas when they are asked what they should do, their responses will be rational, cognitive, thoughtful and cool-headed. These are then the two selves: the 'want self', and the 'should' self.

Loewenstein (1996), in an approach that is to some extent complementary to the previous ones, attributes to 'visceral factors' (hunger, pain, sexual desire, moods and emotions, etc.) the fact that people often act against their self-interest in full knowledge that they are doing so.

There should be no doubt that human beings are sometimes incapable of doing what they think is in their best interest. Doing what one thinks one should do, or the dominion of the should self over the want self, is what Aristotle called moral virtue. In the Aristotelian account, moral virtue is acquired with practice. If that is true, one could expect that the impact of the visceral factors, or the relative importance of the want and should selves, will depend very much on each individual and his/her past history.

The organizational context makes things even more complex. A manager's self-interest may be substantially different from the (otherwise espoused) organizational objective. The manager may say, for instance, that the main goal of the firm is value maximization, and at the same time take actions that destroy long-term value, possibly for short-run benefit. Jensen (2000) and Senge (2000) provide excellent examples of that possibility. Jensen argues that this is the result of "the tendency of human beings to resort to short-term value-destroying actions in the name of value creation" (2000: 50). Indeed, and again according to Jensen, the latest financial scandals have only confirmed this tendency, even to an extreme degree (2002). This analysis leads naturally to the following proposition:

Proposition 7*: Under bounded rationality, a crucial element in trusting another person is whether that person is able to put into practice what he/she thinks would be best for him/herself in the long run, in spite of possibly attractive, short-term results.*

Doing what you think you should do: trust based on character

The analysis in the previous section introduces a new facet of the word 'trust'. Previously, we consistently referred to the decision-making process as if all decisions made by an economic agent were to be immediately implemented with no problem. In the last section we suggested that this may not be so, and that before implementation the decision may change, not because of any new information coming in, or any changes in one's tastes, or any further thoughts on the basic values, but because of lack of control of oneself. In terms of bounded rationality, it can be interpreted that the decision maker's focus of attention shifts to the more immediate, attractive variables, in preference to future variables that are in fact preferable for an individual with unbounded rationality. To implement what one considers to be the 'right' or 'rational' decision, willpower is needed. This is the Aristotelian point of view cited in the previous section.

Different people at different points in time will have different degrees of such willpower. According to Aristotle, this develops through practice. Ordinarily, an individual's willpower to do what he/she considers to be the right thing is 'too little' (as in the Schelling examples), but Benabou and Tirole (2002) have shown how, under some conditions, people sometimes adopt *excessively rigid* rules that result in compulsive behaviors such as miserliness, workaholism, or anorexia. Quite obviously, this 'excess virtue' is also acquired, as in the Aristotelian account, through practice.

Hence, for the trustor A, to assess the probability p that the trustee, B, will 'honor' the trust, he/she has to evaluate in fact two probabilities: the probability p_v, that the values of

the other agent are such that B will choose to 'honor' (possibly in spite of immediate, material rewards for doing the opposite), and the probability p_w that B will in fact have the willpower to put that decision into practice, given that the decision has been made, i.e. (assuming they are independent),

$$p = p_v \cdot p_w$$

Obviously, p_v depends on B's system of values, and the willingness B has to make rational decisions according to that system; while p_w depends on B's willpower, and on the availability of 'attractive' alternative actions to betray A. If there are no immediate, attractive rewards for B to betray A, so that B has no problem honoring A's trust, then p_w will be equal to 1; and the more attractive the rewards available to B for betraying A, the lower p_w will be. All this leads to the following proposition:

Proposition 8*: Under bounded rationality, trust in another person is based on an assessment of three factors: (1) adherence of the other person to a (stable) system of values; (2) that such a system of values includes in some way social goals, or the interests of the other person; and (3) the willpower of the other person to put into practice what she believes she should.*

On the empirical evaluation of the *p's*

There is an important difference between the foundation of trust resting on judgment and competence and the foundation of trust resting on preferences and integrity. The former is entirely empirical: B cannot 'fake' a knowledge that he/she does not have, and once he/she has it, will continue to have it in the future. B, of course, may have to adapt to new situations in the future and learn more, and may be luckier or unluckier in the short run, but the fact remains that the proof of B's competence is in the empirical success in the (average) results of his/her decisions.

The latter, in contrast, cannot be entirely empirical. It is empirical to the extent that all knowledge of the real world comes (obviously) from observation of empirical facts; but there is an important problem associated with assessing someone else's preferences and values: they can be 'faked'. In repetitive decisions, one agent may fake a preference for some variables just to gain someone else's trust, and, then, once this is achieved, betray the other party. In fact, it is rather common for embezzlers to have an immaculate history of honesty, even to excess, until, one day, they betray the trust deposited in them (showing, incidentally, that the trustor was actually vulnerable). So, for both reasons, even if a given person has shown unchanging preferences over a long period of time, one can never be sure that this behavior will continue indefinitely.

This is closely related to a well-known problem in philosophy, the problem of induction. Bertrand Russell (1959) remarked that the fact that for ages we have observed the Sun rising every day does not necessarily imply that it is going to rise tomorrow. The chances that the Sun will rise tomorrow vary greatly with the causal explanation we attribute to its motion. If the sun rises because some giants light a ball of fire at night every night and raise it in the morning, then, if one day they are too tired, or they feel whimsical, they may not light it at all. Yet, this was an explanation believed by some of our ancestors, not too long ago by historical standards. Currently, we believe in the laws of motion, gravitation and the Earth's rotation as causes of the sun rising, and this makes it much more unlikely that the sun will not rise tomorrow. Too many things would have to change. The will of the giants may change much more easily than the motion of enormous bodies and their laws.

Notice that the Russell example involves only physical systems (according to the explanation we accept today, one might add), where there are no reasons for doubting the regularity of the phenomenon. But if induction is a complex issue in the natural sciences, it is even more complex when the system under study is a human being. The parable of the inductivist chicken provided (again) by Russell is very much to the point here:

> "Domestic animals expect food when they see the person who usually feeds them. We know that all these rather crude expectations of uniformity are liable to be misleading. The man who has fed the chicken every day throughout its life at last wrings its neck instead, showing that more refined views as to the uniformity of nature would have been useful to the chicken" (Russell, 1959: 35).

In other words, mere repetition of a given choice by one individual is not a good basis to infer that it will be repeated again. Human beings are purposeful systems, and may have intentions completely different from the ones that seem obvious; and unless there is some understanding of the real reasons why they act, any forecast of their next action may turn out to be completely wrong.

Bounded rationality is again a determining concept. With unbounded rationality, human beings would have unlimited capacity for 'faking' preferences; but they would also have unlimited capacity for discounting for that fact. That is, it would not be too difficult for one person to try to 'trick' the others, but the others would immediately assign a probability to that eventuality and incorporate it in their subjective probability. The success of such a strategy would therefore be in doubt.

With bounded rationality, one agent may try to internalize the value and preference system of the other, partly through previous formal interactions of the same kind, but partly through other means of communication: words, body language, common friends, shared beliefs, attitudes… All these factors are relevant to determine the probability p that B will honor A's trust. The crucial fact, however, is that mere historical repetition of a given alternative in previous, similar situations ('reputation', if by that word we simply mean the other person's track record) is hardly enough. Internalizing the way the other person thinks and her value system is an important element in the assessment of the probability of the other person honoring or betraying trust.

Individualistic-rationalistic approach versus social and cultural approaches

This paper has taken the individualistic-rationalistic approach, which starts from the assumption that trust is rationally based. The analysis in the last section, though, suggests that the social and cultural approach is also needed as a complement. When evaluating the probability of someone else's behavior, social and cultural factors cannot be ignored, mainly in the broader context of not referring only to a specific decision situation, in the 'delegation' setting mentioned before. But we want to show here that, while the individualistic-rationalistic approach is incomplete to deal with the problem of trust, some of the characteristics of trust that are purported to be socially and culturally based are often a social reflection of the rationalistic approach.

Several researchers (e.g., Mayer, Davis & Schoorman, 1995) have emphasized 'propensity to trust' as one of the important characteristics that conditions actual trust, either suggesting that it is each person's personal experience that is at the base of this propensity, or

else pointing out that different cultural backgrounds differ in their propensity to trust (Hofstede, 1980).

As we have seen, Fischoff et al. (1988: 399) argued that habits and traditions can be seen as trial-and-error learning summarized in rules and homilies, and that they provide stereotypic, readily justifiable responses to questions of values. This is particularly relevant in our context, because it means that the bases of trust can be considered to be instrumental from the beginning. Intuitions, rules and traditions provide an initial a priori probability *p* that trust will be honored by the other party, and any subsequent interactions of any kind (verbal or non-verbal communication, real actions between the two individuals...) may modify that probability. But of course, for specific individuals this accumulated learning is transmitted only as a social habit or tradition.

Many of the individualistic-rationalistic models justify social beliefs and attitudes towards trust. Thus, Neilson (1999) develops a model where two agents interact repeatedly in a prisoner's dilemma, and shows that an agent A is willing to do 'costly' favors to another agent B if A expects to receive favors in return in the future. His approach is quite clearly individualistic-rationalistic; but, of course, creating the social climate where one expects reciprocity in doing favors makes it easier for cooperation to exist. Along similar lines, Spagnolo (1999) shows how workers have an incentive to cooperate if the probability that the other party will cooperate is large enough. Valley, Moag and Bazerman (1998) show how, with asymmetric information, the communication medium (which obviously is a social creation) affects the distribution of outcomes, reflecting different degrees of truth-telling and trust across the media. Tullock (1999) changes the usual conditions of experiments on the prisoner's dilemma (i.e., does not pre-select contestants, does not prevent them from communicating, and does not change partners in the middle), and gets a very high degree of cooperation, in contrast with what happens under the usual conditions.

Thus, many unconscious habits may have their origins in rational attitudes. Dasgupta provides a good description of many of these factors from the rational perspective:

> "We form an opinion on the basis of his background, the opportunities he has faced, the courses of action he has taken, and so forth. Our opinion is thus based partly on the theory we hold of the effect of culture, class membership, family line, and the like on a person's motivation (his disposition) and hence his behavior" (1988: 54).

But notice that two kinds of elements enter into this description. Background, culture, class membership, family line, and so on may be considered (paradoxically perhaps) elements of the individualistic-rationalistic approach. They are elements that may indicate the kind of person the hypothetical trustee is, and what can be specifically expected from such a person: first-hand experiences, reputation, track record, and so on.

In contrast, "the theory we hold of the effect of...." is clearly a cultural creation of the social group to which the would-be trustor belongs. Obviously, though, those social influences do not exhaust the explanation of one individual's propensity to trust, which by necessity must include personal factors.

The line between the individualistic-rationalistic point of view and the influence that the social environment exerts upon individuals is difficult to draw. Probability *p* is partly determined by the social background of the individual, and partly by the individual's direct experience. But social, cultural and relational aspects of trust have a background of instrumentality behind them.

Tyler and Degoey (1996: 339) analyze in some depth the reasons why the instrumental view might not be enough to explain the phenomenon of trust. They give three arguments. First, if trust were merely instrumental, "people will care about trustworthiness when they are dependent on the organization or vulnerable to harm". Instead, what they found is that trustworthiness is central when people have "a personal connection with the authorities or identify with the organization". While this claim may of course be true, a "personal connection" with the authorities, or "identification" with the organization requires some knowledge of the authorities' value system, on which that trust can be based, according to this paper's analysis, on an instrumental basis.

Second, one would expect from the instrumental model that trust would be "linked to satisfaction with the authority's decisions"; while if it is relational, it should be "linked to judgments about the neutrality of authorities and the degree to which these authorities treat their subordinates with dignity and respect". Again, under a rational-instrumental approach it is perfectly possible to argue that the value system of the authorities is at the root of the subordinates's trust in them.

Finally, if trust is instrumental in character, "judgments about the competence of authorities should be more strongly linked to people's willingness to accept an authority's decision than judgments about the benevolence of authorities". The way the problem of trust has been analyzed in this paper, that claim would have to be denied. To the extent that the willingness to accept an authority's decision is related to trust in that authority, that trust might in fact be based partly on competence, and partly on value systems, a particular case of which is benevolence.

In summary, and as I have stated already, some of the characteristics of trust that are often assumed to be socially and culturally based, may also be a social reflection of a rationalistic attitude.

RECAPITULATION AND CONCLUSIONS

Trust is a complex subject. Elusive, as was recognized from the beginning, and with many facets. It can have very different meanings, which we have tried to explore analytically in this paper, and it is time now to recapitulate and see where everything stands.

A first conclusion is that if rationality is unbounded and information is symmetrical among the two agents involved, "trust" cannot exist in any meaningful way. Under these conditions, we are led to what James called "the paradox of trust": the only possible way to get one person to trust another is by changing the payoffs and making trust unnecessary. Formally, in a one-time interaction, this essentially means making an enforceable explicit contract where the two parties commit themselves to the action that leads to Pareto optimality.

The possible existence of utility for variables of a qualitative, intangible nature does not substantially change the conclusion, provided rationality is unbounded, and therefore the two agents know perfectly what they want, and make no mistakes. What matters, then, is total utility to the two agents. If they both value the outcomes of the branch of the "trust-honor" tree higher than any other branch, there will be no need for trust.

Trust can be meaningful, though, even in the absence of any conflict of interest, to the extent that there is asymmetrical information about outcomes and the trustor "decentralizes"

the decision because of the specialized information of the trustee, "trusting" that his/her information is 'better'. What in the literature has been called "trust in competence" is related to that meaning. This would be the second conclusion.

Bounded rationality, which in this article's context essentially stresses the point that agents do not have full knowledge of their own preferences and values, completely changes the meaning of trust, and the reasons for its existence. Under bounded rationality, preferences are organized in 'value systems', but decisions made may or may not be consistent with them. A person's willingness to act according to the 'higher values' (typically few in number) is one possible reason to 'trust' that the person will follow some course of action that may be in that person's own best interest, but perhaps not the most attractive in terms of the immediate variables of both effort and results. 'Trust', then, means the belief by the trustor that the trustee will make the decision according to his/her real value system, even if some immediate variables push her in the opposite direction. That is our third conclusion.

The fourth conclusion is rather intuitive. If, besides being consistent with a value system, some of the trustee's 'higher values' are non-selfish, and so include, say, truthfulness, friendship, social welfare, etc.., they provide a better basis for trust, i.e., the trustor may believe that the trustee will not take advantage of his/her vulnerabilities. Here, in contrast with the previous situation, where values were not necessarily non-selfish, the concept of trust may go beyond specific situations and extend to a class of decisions. This is, therefore, the concept that provides a foundation for taking some risks in situations of decentralization of authority, giving power to the trustee for a certain type of decisions.

Finally, whatever the actual preferences and values are, the trustee's action depends on his/her capacity to actually put into practice what he/she thinks is good according to his/her own system of values; and therefore, trusting someone means trusting his/her capacity to do precisely that.

Appendix 1

Suppose that the two agents share equally the result of two possible alternative actions, a1 and a2, which depends on a state variable that can take two values, 0 and 1. The two agents are risk-neutral, but have different information about the state variable. Agent A is completely ignorant about that variable, and attributes equal probability to both states. Agent B is an 'expert' and has a different probability assessment (say, 70% for state 0, and 30% for state 1). The payoff matrix for each agent is shown as Table 1.

Table 1
Payoff matrix

	Action a_1	Action a_2
State 0	10	8
State 1	–5	–2

Under A's probability assessment, action a_2 is preferable (it has an expected value of 3, while that of a1 is only 2.5); while under B's probability assessment, a_1 has an expected value of 5.5, and a_2 of 5, making the first action more desirable. In this uncertainty context, 'A trusts B' may have a meaning completely unrelated to the difference in results for the two agents: A may trust B's information better than his/her own.

Appendix 2

In the context of the data in Table 2, the obvious choice for B, if trusted, is to honor A's trust (his/her utility is greater); and for A, knowing this, the probability that B will honor his/her trust is 1, and his/her obvious choice is "trust". If the utility to B of betraying A when trusted were instead (say) 200, then his/her obvious choice would be to betray; and A, knowing this, would assign a probability of 0 to B honoring his/her trust.

Table 2
Utility for non-measurable variables

	Monetary rewards				**Total utility**			
	B honors		B betrays		B honors		B betrays	
	To A	To B	To A	To B	To A	To B	To A	To B
A trusts	10	10	–5	20	150	150	100	100
A doesn't	0	0	0	0	0	0	0	0

Appendix 3

Assume there is certainty, and that the monetary results are as shown on the left-hand side of Table 3. Let us further assume that the utility functions for monetary rewards are logarithmic, namely, that for each agent, his/her selfish utility is

$$u(x_i) = \log (5+x_i),$$

while the altruistic utility is

$$v(x_i) = k_i \log (5+x_j),$$

and additive to the former, so that total utility is

$$w_i(.) = \log (5+xi) + k_i \log(5+x_j)$$

For a completely selfish individual, k=0; and the larger the k, the more 'benevolent' that individual will be towards the other individual. Negative values of k would, of course, represent "malevolent" individuals who dislike others being happy.

Then, taking the numbers in our example, the corresponding utilities are as shown in Table 3. We can see there that both players have "trust" as a dominant strategy now, and it is the one that leads to a Pareto-optimal solution from a purely monetary point of view. A's utility, in the event that B 'betrays', is minus infinity, which might be interpreted to mean that A would die of starvation, or live in total deprivation. B is unwilling to be party to that eventuality, and so is willing to sacrifice part of his/her own wealth to prevent it from happening.

Table 3
Non-selfish utilities

	Monetary rewards				**Total utility**			
	B honors		B betrays		B honors		B betrays	
	To A	To B	To A	To B	To A	To B	To A	To B
A trusts	10	10	–5	20	5.42	5.42	–∞	–∞
A doesn't	0	0	0	0	3.22	3.22	3.22	3.22

Appendix 4

Psychological states associated with not knowing what you want

Having no opinion
- Not realizing it
- Realizing it
  - Living without one
  - Trying to form one

Having an incoherent opinion
- Not realizing it
- Realizing it
  - Living with incoherence
  - Trying to form a coherent opinion

Having a coherent opinion
- Accessing it properly
- Accessing only a part of it
- Accessing something else

References

Aghion, Ph. & Tirole, J. 1997. Formal and Real Authority in Organizations. ***Journal of Political Economy***, 105 (1): 1-29.

Arrow, K.J. 1974. ***The Limits of Organization***. New York: Norton and Company.

Barnard, Ch. I. 1938. ***The Functions of the Executive***. Cambridge, Mass.: Harvard University Press.

Barney, J.B. & Hansen, M.H. 1997. Trustworthiness as a Source of Competitive Advantage. In H.T. O'Neal and M. Ghertman (Eds.), ***Strategy, Structure and Style***: 5-22. New York: JohnWiley and Sons.

Benabou, R. & Tirole, J. 2002. ***Willpower and Personal Rules***, CEPR Discussion Paper No. 3143, January.

Bazerman, M.H., Tenbrunsel, A.E. & Wade-Benzoni, K. 1998. Negotiating with yourself and losing. ***Academy of Management Review***, 23 (2): 225-241.

Casadesus-Masanell, R. 2004. Trust in Agency. ***Journal of Economics and Management Strategy***, 13(3): 375-404.

Coleman, J.S. 1990. ***Foundations of Social Theory***. Cambridge, Mass: Harvard University Press.

Dasgupta, P. 1988. Trust as a commodity. In D. Gambetta (Ed.), ***Trust: Making and Breaking Cooperative Relations***: 49-72. Oxford: Basil Blackwell.

Deutsch, Morton, 1962. Cooperation and Trust: Some Theoretical Notes. In R. Jones Marshall (Ed.), ***Nebraska Symposium on Motivation***: 275-319. University of Nebraska.

Fischhoff, B., Slovic P. & Lichtenstein, S. 1988. Knowing what you want: measuring labile values. In D.E. Bell, H. Raiffa & A. Tversky (Eds.), ***Decision-making. Descriptive, normative and prescriptive interactions***: 398-421. Cambridge, UK: Cambridge University Press.

Fukuyama, F. 1995. ***Trust: Social Virtues and the Creation of Prosperity***. New York: The Free Press.

Gabarro, J. 1978. The development of trust, influence and expectations. In A.G. Athos & J. Gabarro (Eds.), ***Interpersonal Behavior: Communication and Understanding in Relationships***: 290-303. Englewood Cliffs, NJ: Prentice-Hall.

Gambetta D. 1988. Can We Trust Trust?. In D. Gambetta (Ed.), ***Trust: Making and Breaking Cooperative Relations***: 212-237. Oxford: Basil Blackwell.

Giddens, A. 1990. ***The Consequences of Modernity***. Stanford, Cal.: Stanford University Press.

Granovetter, M. 1985. Economic action and social structure: the problem of embeddedness. ***American Journal of Sociology***, 91: 481-510.

Hirschman, A.O. 1984. Three easy ways of complicating economic discourse. ***American Economic Review***, 74 (2): 89-96.

Hofstede, G. 1980. Motivation, Leadership and Organization: Do American Theories Apply Abroad? ***Organizational Dynamics***, 9(1): 42-63.

James, H.S. 2002. The trust paradox: a survey of economic inquiries into the nature of trust and trustworthiness. ***Journal of Economic Behavior and Organization***, 47: 291-307.

Jensen, M. 2000. Value maximization, stakeholder theory and the corporate objective function. In M. Beer and N. Nohria (Eds.), ***Breaking the Code of Change***: 35-57. Boston, Mass.: Harvard University Press.

Jensen, M. 2002. ***Just Say "No" to Wall Street.*** Negotiation, Organization and Markets Unit, Harvard Business School, Working Paper No. 02-01.

Kahneman, D., Knetsch, J.L. & Thaler, R.H. 1986. Fairness and the assumptions of economics. ***Journal of Business***, 59: 285-300.

Korczynski, M. 2000. The political economy of trust. ***Journal of Management Studies***, 37(1): 1-21.

Kramer, R.M. & Tyler, T.R. 1996. Whither Trust?. In R.M. Kramer & T.R. Tyler (Eds.), ***Trust in Organizations***: 1-15. Thousand Oaks: Sage Publications.

Kreps, D.M. 1990. Corporate Culture. In J.E. Alt & K.A. Shepsle (Eds.), ***Perspectives on Positive Political Economy***: 90-143. New York: Cambridge University Press.

Loewenstein, G. 1996. Out of Control: Visceral Influences on Behavior. ***Organizational Behavior and Human Decision Processes***, 65 (2): 272-92.

March, J. 1987. Ambiguity and Accounting: The Elusive Link Between Information and Decision-Making. ***Accounting, Organizations and Society***, 12: 153-168.

Mayer, R., Davis, J. & Schoorman, F. 1995. An Integrative Model of Organizational Trust. ***Academy of Management Review***, 20 (3): 709-734.

Misztal, B. 1998. ***Trust in Modern Societies***. Cambridge, UK: Polity Press.

Neilson, W.S. 1999. The economics of favors. ***Journal of Economic Behavior and Organizations***, 39: 387-397.

O'Donoghue, Ted, and Matthew Rabin, 1999. Doing it Now or Later. ***American Economic Review***, vol. 89 (1): 103-124.

Portales, C., Ricart, J.E. & Rosanas, J.M. 1998. Understanding Trust to Build Strong Relationships in Organizations. In M. Hitt, J.E. Ricart & R.D. Nixon (Eds.), ***Managing Strategically in an Interconnected World***. New York: John Wiley and Sons.

Rokeach, M. 1973. ***The Nature of Human Values***. New York: The Free Press.

Russell, Bertrand, 1959. ***The Problems of Philosophy***. Oxford: Oxford University Press.

Schelling, T. 1978. Egonomics or the Art of Self-Management. ***American Economic Review***, 68 (2): 290-294

Schelling, Thomas, 1984. Self-Command in Practice, in Policy and in the Theory of Rational Choice, ***American Economic Review***, 74 (2): 1-11.

Seligman, A. 1997. ***The Problem of Trust***. Princeton, N.J.: Princeton University Press.

Selten, R. 1999. ***What is bounded rationality?*** Discussion Paper B-454, Rheinische Friedrich-Wilhelms Universität, Bonn.

Senge, P. 2000. The Puzzles and Paradoxes of How Living Companies Create Wealth. Why Single-Valued Objective Functions Are Not Quite Enough. In M. Beer and N. Nohria (Eds.), ***Breaking the Code of Change***: 59-81. Boston, Mass.: Harvard University Press.

Shapiro, D.L. Sheppard, B.H. & Cheraskin, L. 1992. Business on a handshake. ***Negotiation Journal***, 8: 365-378.

Simon, H.A. 1983. ***Reason in Human Affairs***. Oxford: Basil Blackwell.

Simon, H.A. 1987. Making Management Decisions: the Role of Intuition and Emotion. ***Academy of Management Executiv***e, 1 :57-64.

Simon, H.A. 1991. Organizations and Markets. ***Journal of Economic Perspectives***, 5 (2): 25-44.

Simon, H.A. 1997. ***Administrative Behavior***, fourth edition. New York: The Free Press.

Spagnolo, G. 1999. Social relations and cooperation in organizations. ***Journal of Economic Behavior and Organizations***, 38: 1-25.

Thaler, R. & Shefrin, H.M. 1977. An Economic Theory of Self-Control. ***Journal of Political Economy***, 89 (21): 392-406.

Tullock, G. 1999. Non-prisoner's dilemma. ***Journal of Economic Behavior and Organizations***, 39: 455-458.

Tyler, T.R. & Degoey, P. 1996. Trust in organizational authorities: the influence of motive attributions on willingness to accept decisions. In R.M. Kramer & T.R. Tyler (Eds.), ***Trust in Organizations***: 331-356. Thousand Oaks: Sage Publications.

Valley, K.L., Moag, J. & Bazerman, M.H. 1998. A matter of trust: Effects of communication on the efficiency and distribution of outcomes, ***Journal of Economic Behavior and Organizations***, 34: 211-238.

Webster's Seventh New Collegiate Dictionary, 1972. Springfield, Mass.: G&C Merriam Co.

Williamson, O.E. 1985. ***The Economic Institutions of Capitalism***. New York: The Free Press.

Williamson, O.E. 1993. Calculativeness, trust and economic organization. ***Journal of Law and Economics***, 34: 453-502.

Zand, Dale E. 1972. Trust and Managerial Problem Solving. ***Administrative Science Quarterly***, 17: 229-239.

Zucker, L. 1986. Production of trust: institutional sources of economic structure. In Staw, B.M & Cummings, L.L. (Eds.), ***Research in Organizational Behavior***, 8: 53-111.