CentER

Discussion Paper

No. 2006–37

HUMAN NATURE IN THE ADAPTATION OF TRUST

By Bart Nooteboom

April 2006

ISSN 0924-7815

Tilburg University

HUMAN NATURE IN THE ADAPTATION OF TRUST

Bart Nooteboom
Tilburg University
Adelheidstraat 82, 2595 EE the Hague, the Netherands
Phone: +31703478605
e-mail: b.nooteboom@uvt.nl

Abstract

This chapter pleads for more inspiration from human nature, in agent-based modeling. As an illustration of an effort in that direction, it summarizes and discusses an agent-based model of the build-up and adaptation of trust between multiple producers and suppliers. The central question is whether, and under what conditions, trust and loyalty are viable in markets. While the model incorporates some well known behavioural phenomena from the trust literature, more extended modeling of human nature is called for. The chapter explores a line of further research on the basis of notions of mental framing and frame switching on the basis of relational signaling, derived from social psychology.

JEL code: A14, D01, D64, L14, L24, Z13

Key words: Trust, transaction costs, buyer-supplier relationships, social psychology

Introduction

For the object of study I choose trust, for several reasons. First, if anything is human, it is (dis)trust. Second, if anything is subject to adaptation, it is trust, in its build-up and breakdown, and as both the basis for a relationship and its outcome. Third, trust forms an important issue in economics, and in behavioural science more widely. Trust is needed to limit transaction costs and costs of contracting and control. In the literature on transaction costs and inter-firm relations there has been a debate whether trust can exist in markets, under pressures of competition. Agent-based simulation seems an appropriate tool for experimentation, to investigate under what conditions trust is viable in markets.

Many attempts have been made at agent-based modeling of trust and related issues. The purpose of trust models varies widely. Generally, they study emergent properties of complex interaction that would be hard or impossible to tackle analytically. Some study the effectiveness of sanctions and/or reputation mechanisms and agencies to support them, e.g. in information systems or supply chains (Zacharia et al., 1999; Meijer & Verwaart, 2005; Diekmann & Przepiorka,

2005), or in artificial societies (Younger, 2005). Some study self-organization, e.g. in the internalization of externalities in a common pool resource (Pahl-Wost & Ebenhöh, 2004), the emergence of leadership in open-source communities (Muller, 2003), or the emergence of cooperative social action (Brichoux & Johnson, 2002). Others investigate the working of decision heuristics (Pahl-Wost & Ebenhöh, 2004; Marsella et al., 2004).

The general set-up is that of multiple agents who can profit from each other but are uncertain about the quality or competence that is offered, sometimes allowing for multiple dimensions of quality, and dependencies between them (Maximilien & Singh, 2005). Other studies focus on the benevolence or intentions of agents, i.e. absence of cheating, in free-ridership, defection or expropriation of knowledge or other resources, and many look at both competence and intentions (Castelfranchi & Falcone, 1999; Pahl-Wost & Ebenhöh, 2004; Breban, 2002; Muller, 2003; Gans et al., 2001). This is line with the distinction made in the trust literature between competence trust and intentional trust (see e.g. Nooteboom, 2002).

Mostly, agents are oriented only towards their self-interest, such as maximum profit, but some studies also allow for fairness and equity as objectives or dimensions of value (Pahl-Wost & Ebenhöh, 2004). Mostly, trust is measured as a number beween zero and one, and, following Gambetta (1988), is often interpreted as a subjective probability that goals will be achieved or no harm will be done. Mostly, conduct is individual, but sometimes allowance is made for coalitions (Breban, 2002).

Few studies of defection explicitly model both sides of the coin: the expectation of defection by others (trust) and one´s own inclination to defect (trustworthiness). Also, most studies treat trust as of purely extrinsic value, in the achievement of profit, and do not include the possible intrinsic value of trust. Notable exceptions are Pahl-Wost & Ebenhöh (2004) and Marsella et al. (2004).

Trust is generally updated on the basis of experience, sometimes only one's own experience in interaction, sometimes (also) on the basis of reputation mechanisms, sometimes with the services of some ´tracing agency´ (Zacharia et al., 1999; Meijer & Verwaart, 2005; Diekman & Przepiorka, 2005). Few studies are based on an explicit inference of competence or intentions, and even fewer studies explicitly model the decision heuristics used. Exceptions here also are Pahl-Wost & Ebenhöh (2004) and, with great psychological sophistication, Marsella et al. (2004). Those studies will be considered in more detail later. A key question is whether agents have 'a theory of mind' on the basis of which they attribute competencies and intentions to others.

While most studies model trust as adaptive, in the sense that it develops as a function of private or public experience, there is very little study, as far as I know, of adaptiveness of the importance attached to trust relative to profit, and of the adaptiveness of one´s own trustworthiness or inclination to defect.

In this chapter, by way of illustration, a model is discussed with some of these features. It focuses on intentional trust, in terms of loyalty or defection, based on private experience (no reputation effects). Trust is adapted on the basis of observed defection, but only with simple reinforcement, without theory of mind and explicit decision heuristics. Next to trust it includes own trustworthiness, i.e. inclination to

defect. Trustworthiness and the importance attached to trust are both adaptive, as a function of experience.
The central purpose of the study is theoretical: to investigate whether the claim of transaction cost economics that trust cannot survive under competition (Williamson, 1993) is correct. Under what conditions, if at all, are trust and trustworthiness viable in markets where the performance criterion is purely profit? The analysis is conducted in the context of transaction relations between multiple buyers and suppliers, which is the classical setting for the analysis of transaction costs. Thus, the present chapter is related to other sections of the present volume, on 'Industrial structures and innovation' and 'Supply chain management'.
This chapter proceeds as follows. First, it summarizes this example of an agent-based computational model of trust. Second, it explores possibilities to proceed further in this direction, in an attempt to bring more human nature into the modeling of trust, in the employment of decision making heuristics offered by social psychology.

A Model of Adaptive Trust

Trust

Trustworthiness may be based on self-interest, but also on benevolence, based on solidarity or loyalty. This is related to two different definitions of trust. According to one definition, trust entails vulnerability of the trustor to possibly harmful actions of the trustee, with the expectation that, for whatever reason, no great harm will be done. The reasons for this expectation may include control or deterrence, in which the trustee refrains from opportunism either because he has no opportunity for it, due to contractual or hierarchical constraints, or no incentives for it, since he is dependent on the trustor or wishes to protect his reputation. For this general notion, which includes safeguards on the basis of control, Nooteboom (2002) proposed not to use the term 'trust' but the more general term of 'reliance'. Reasons for trustworthiness may also include motives that go beyond (narrow) self-interest, such as the wish to behave appropriately, according to social or moral norms or values, or empathy or identification with the trustor, in combination with feelings of sympathy, friendship or solidarity (MacAllister, 1995; Lewicki & Bunker, 1996). This is what people mostly mean by the term 'trust'.

Is trust viable in markets?

I will summarize and discuss a model of the emergence and adaptation of trust published by Klos & Nooteboom (2001). The purpose of the model was to develop a tool for assessing the viability of trust, in the sense of benevolence, between firms in markets. That is a much-debated issue (for a survey, see Nooteboom, 2002). Economics, in particular transaction cost economics (TCE), doubts the viability of benevolence, on the argument that under competition, in markets, firms are under pressure to utilize any opportunistic opportunity for profit (Williamson, 1993). However, especially under the uncertainty and volatility of innovation reliance on the basis of control, such as complete contracts, but also reputation mechanisms, are infeasible or unreliable, so that especially

there benevolence is needed as a basis for governance, as a substitute or complement for necessarily incomplete contracts (Nooteboom, 1999, 2004) and reputation mechanisms. Thus, it is of some theoretical and practical importance to investigate whether, or when, benevolence may be viable. I propose that benevolence, going beyond calculative self-interest, can exist in markets but is nevertheless subject to circumstances, such as pressures of survival, depending on intensity of competition and the achievement of profit (Pettit, 1995), and experience. The purpose of the model is to explore these circumstances.
To serve its purpose, the model should incorporate essential elements of TCE logic. TCE proposes that people organize to reduce transaction costs, depending on conditions of uncertainty and specific investments, which yield switching costs and a resulting risk of 'hold-up'. The model employs TCE logic, but also deviates from TCE in two fundamental respects. First, while TCE assumes that optimal forms of organization will arise, yielding maximum efficiency, that is problematic. The making and breaking of relations between multiple agents with adaptive knowledge and preferences may yield complexities and path-dependencies that preclude the achievement of maximum efficiency. Even if all agents can in principle access all relevant partners, and have relevant knowledge about them, actual access depends on competition for access, and on unpredictable patterns of making and breaking relations among multiple agents. Second, while TCE assumes that reliable knowledge about loyalty or trustworthiness is impossible (Williamson, 1975), so that opportunism must be assumed, it is postulated here that to some extent trust is feasible, by inference from observed behaviour.
The methodology of Agent Based Computational Economics (ACE) is well suited to model complexities of multiple interactions, and to see to what extent theoretical benchmarks of maximum efficiency can in reality be achieved. It enables us to take a process approach to trust (Zand, 1972; Zucker, 1986; Smith Ring &van de Ven, 1994; Gulati, 1995), by modeling the adaptation of trust and trustworthiness in the light of experience in interaction.

The model

In the model, buyers and suppliers are matched on the basis of preferences that are based on both trust and potential profitability, where trust can also have intrinsic value. In this matching, depending on their preferences agents make, continue or break transaction relations. Trust is based on observed loyalty of partners, i.e. absence of switching to a different partner. In line with industrial economics, profit is a function of product differentiation (which increases profit margin), economy of scale from specialization, and learning by cooperation in ongoing relations. Use is made of the notion (from TCE) of specific investments in relationships. Those have value only within the relationship, and thus would have to be made anew when switching to a different partner. Specific investments yield more differentiated products, with a higher profit margin. Economy of scale yields an incentive for buyers to switch to a supplier who supplies to multiple buyers, which yields a bias towards opportunism, in breaking relations with smaller suppliers. However, this can only be done for activities that are based on general-purpose assets, not relation-specific investments for specialty products.

The percentage of specialty products is assumed to be equal to the percentage of specific investments, as a parameter of the model that can be set. The specialty part, which is relation-specific, yields higher profit, and is also subject to learning by cooperation, as a function of an ongoing relation. Thereby, it yields switching costs, and thus yields a bias towards loyalty.
In sum, the model combines the essential features of TCE: opportunism by defection, specific investments, economy of scale for non-specific investments, and switching costs. However, the model adds the possibility of trust as a determinant of preference, next to potential profit.
In the model, agents are adaptive in three ways. In the preference function, specified in an appendix, the relative weights of potential profit and trust are adaptive, as a function of realized profit. In this way, agents can learn to attach more or less weight to trust, relative to potential profit. Agents adapt their trust in a partner as a function of his loyalty, exhibited by his continuation of the relationship. As a relation lasts, trust increases incrementally, but with decreasing returns, and it drops discontinuously when defection occurs. Agents also adapt their own trustworthiness, modeled as a threshold of exit from a relation, on the basis of realized profit. Agents only defect, in switching, when incremental preference exceeds the threshold. This models the idea that while agents maybe loyal, that has its limits. Thus, agents can learn to become more or less trustworthy in the sense of being loyal.
Note that adaptation of both the weight attached to trust and the threshold of defection occurs on the basis of realized profit. This biases the model in favor of Williamson's (1993) claim that trust cannot survive in markets. In the model, trust and trustworthiness can only emerge when they enhance realized profit. The model allows us to explore under what conditions, in terms of parameter settings, trust and loyalty increase, or are stable, i.e. when they are conducive to profit, and hence viable in markets.
Starting values of agent-related parameters, such as initial trust, threshold of defection, and weight attached to trust, can be set for each agent separately. This allows us to model initially high or low trust societies, in setting parameters accordingly for all or most agents, or to model high trust agents in low trust societies, and vice versa, to study whether and when trust is viable, or is pushed out by opportunism. Other, non agent-related parameters, such as the percentage of product differentiation and specific assets, strength of economy of scale, strength of learning by cooperation, speed with which trust increases with duration of a relation, number of buyers, number of suppliers, and number of time steps in a run, are fixed per experiment.
In sum, the model is set up to experiment with conditions for trust to grow or decline, as a function of realized profit, depending on trade-offs between advantages of defection (for economy of scale) and advantages of loyalty (in learning by doing in an ongoing relationship). Further technical details of the model are specified in Appendix A.

Simulation results

Initial expectations were as follows:

- In interactions between multiple, adaptive agents, maximum efficiency is seldom achieved, due to unforeseeable complexities of interaction
- In conformance with TCE, in the absence of trust outsourcing occurs only at low levels of asset specificity
- High trust levels yield higher levels of outsourcing at all levels of asset specificity
- Under a wide range of parameter settings, high trust levels are sustainable in markets
- The choice between an opportunistic switching strategy and loyalty depends on the relative strength of scale effects and learning by cooperation

All these expectations are borne out by recent simulation experiments (Gorobets & Nooteboom, 2005). Of course, simulation is not equivalent to empirical testing. The test is virtual rather than real. It has only been shown that under certain parameter settings emergent properties of interaction satisfy theoretical expectations. The significance of this depends on how reasonable the assumptions in the model and the parameter settings are considered to be.

The overall outcome is that both trust and opportunism can be profitable, but they go for different strategies. This suggests that there may be different individual agents or communities, going for different strategies, of switching or of loyalty, which settle down in their own self-sustaining systems. If we compare across the different settings of high, medium and low initial trust, under different conditions concerning the strength of scale effect relative to learning by cooperation, and concerning initial weight attached to trust and initial thresholds of defection, profit declines more often than it increases, as we go from high to low trust. Further details are given in Appendix B.

The following paradox emerges from the analysis. Potential profit from learning by cooperation is highest for the highest level of product differentiation, but precisely then, when trust is low buyers prefer to make rather than buy, and thereby forego the opportunities for learning by cooperation. When buyers focus on profitability rather than trust, profit from economy of scale is instantaneous while learning by cooperation is slow, and the potential for economy of scale is low at high levels of differentiation. Thus, under low trust and low weight attached to it, buyers lock themselves out from the advantages of collaboration. When they outsource, it is mostly at low levels of differentiation, when learning by cooperation yields only modest returns, but then they learn to appreciate its accumulation in lasting relationships. They wind up in outsourcing at high differentiation only ‘by mistake’, then learn to appreciate it, and once learning by doing gets under way, a focus on profit keeps them in the relationship. In time, as profit turns out to be consistent with loyalty and trust, they learn to attach more weight to them. This illustrates a principle noted before, in the trust literature. As a default, i.e. a stance taken until reasons for an alternative stance appear (Minsky, 1975), trust is to be preferred to distrust. Excess trust can be corrected on the basis of experience with untrustworthy partners, while distrust prevents one from engaging in collaboration to learn that partners are in fact trustworthy, if that is the case.

More Human Nature

Discussion of the model

In the model, human nature is modeled to some extent. Trust is reinforced, incrementally, by observed loyalty, and drops discontinuously in case of observed disloyalty. In evaluating an actual or potential relationship, agents consider both potential profitability and trust they have, and the weight attached to the one relative to the other is adapted on the basis of experience, in the form of past profit. Similar adaptation applies to their own trustworthiness (absence of defection).
However, modeling of cognition and decision-making is still primitive in that:

- The rationality of agents is bounded in that they do not take into account opportunities that they have no own experience with, but that are observable. In particular, non-trusting agents who rob themselves of the opportunity to learn that collaboration and loyalty may be profitable do not learn from observing such profit of more trusting competitors.
- In assessing trustworthiness from observed behaviour, agents are myopic, looking only at their own experience with the agent. In other words, the model does not contain a reputation mechanism and gossip.
- Adaptation is highly automatic, in combination with random shifts. There is no modeling of processes of inference and decision-making, and of the emotions involved. The model does not incorporate a theory of mind.

The first two shortcomings can be repaired without a fundamental shift of model design, by including spillovers of experience in a reputation mechanism, where profits and experiences of loyalty of (some or all) other agents contribute to one's adaptation. The third shortcoming is much more fundamental, since it requires the modeling of social-cognitive processes. Options for doing this are explored in the remainder of this chapter.

More human nature from other studies

Two studies I found stand out in their dealing with human nature in processes of inference and decision making. Pahl-Wost & Ebenhöh (2004) emphasize a human approach in terms of decision heuristics and mental that agents select from, such as a frame oriented towards cooperation or towards maximizing, as well as switching between frames as a function of experience. They recognize a range of relevant mental categories in cooperative behaviour: cooperativeness, fairness (concerning others and concerning me), conformity, reciprocity (positive and negative, in retribution), risk aversion, commitment, and trustworthiness (not being opportunistic). A large and necessary step in the modeling of agents is to equip them with a theory of mind, i.e. a basis for inferring competencies and intentions of other agents, as a basis for their decision making. This route of taking decision heuristics known from social psychology is also taken, with impressive sophistication, by Marsella et al. (2004), in their development of virtual agents. This modeling, of beliefs, influence and belief change, is intended as a training device, e.g. for teachers to learn how to deal with bullies in the classroom.
I am confident that this is the way to go, for some applications at least. Reich (2004) pleads for the use of formal logic in the analysis of reactions to actions, and anticipated

reactions to that, on the basis of decision rules. That is no doubt valid, but a socio-cognitive theory is needed to specify those rules. Below, I elaborate some further ideas to proceed along this line of bringing in more human nature, also using insights from social psychology.

Deliberative and automatic response

The trust literature recognizes a duality of rational and automatic response. In social psychology, Esser (2005) also recognized rational deliberation and automatic response as two modes of 'information processing'. However, the non-deliberative or automatic mode seems to split into two different forms: unemotional routine and emotion-laden impulse, out of faith, friendship, suspicion, in a leap of faith or a plunge of fear. Emotions, which determine 'availability' to the mind, as social psychologists call it, may generate impulsive behaviour and may trigger a break of routinized behaviour. A question then is whether the latter automatically triggers an automatic response, or whether an emotionally triggered break with routine can lead on to a rational deliberation of response. For that, the emotion would have to be somehow neutralized, controlled, supplemented, or transformed for the sake of deliberation. In the build-up and breakdown of trust this is of particular importance in view of the indeterminacy of causation. Expectations may be disappointed due to mishaps, lack of competence or opportunism, and it is often not clear which is the case.
If a relationship has been going well for some time, trust and trustworthiness may be taken for granted, in routinized behaviour. A jolt of fear from exceptional events may be needed to break out of the routine, but in view of the causal ambiguity of what went wrong, one may need to give the trustee the benefit of the doubt, allowing for mishaps or lack of competence, rather than jumping to the conclusion of opportunism. When does this happen and when not?
In the trust literature, it has been proposed that as a relationship develops, at some point reliance (whether it is based on control or trust) is based on cognition, i.e. on knowledge concerning the intentions and capabilities of a trustee. Subsequently, actors may develop 'empathy', i.e. understanding of how a partner feels and thinks, and next partners may develop 'identification', i.e. they see their fortunes as connected and they start to feel and think alike (McAllister, 1995; Lewicki and Bunker, 1996). As noted by Luhmann (1980), when people start to cooperate, they get the chance to adopt each other's perspectives. In empathy trust may be associated with feelings of solidarity and in identification with feelings of friendship. In going from knowledge based trust to empathy and identification based trust, behaviour appears to become less deliberative and more automatic, due to both emotions and routinization.

Mental framing

The question now is how we can further clarify the trust process, in terms of how people think and judge, making and adapting interpretations and choices of action, in a fashion that is amenable, at least in principle, to inclusion in an agent-based model.
For this, I employ the notion of mental 'framing', adopted from sociology and social psychology (Lindenberg 2000, 2003; Esser 2005). According to Esser, a mental frame is

an ‘situation defining orientation’ that consists of ‘.. two simultaneously occurring selections: the selection of a mental model of the situation on the one hand and that of the mode of information processing in the further selection of action’(Esser 2005: 95, present author’s translation from the German). Thus, a mental frame is also associated with action scripts of response appropriate for enacting the frame. For mental frames, Lindenberg (2003) recognized three: ‘acting appropriately’ (AA), also called the ‘solidarity frame’ (Wittek, 1999), ‘guarding one’s resources’ (GR), to ensure survival, and a ‘hedonic frame’(H), where one gives in to temptations for gratifying the senses. These three frames are adopted here because they align closely with the distinction, in the trust literature, between ‘benevolence’ and ‘opportunism’, with the latter including both pressures of survival, which seems close to ‘guarding one’s resources’, and vulnerability to temptation when it presents itself, which seems close to the ‘hedonic frame’. The frames may support or oppose each other, and while at any moment one frame is ‘salient’, in determining behaviour, conditions may trigger a switch to an alternative frame.

If frames serve to both ‘define a situation’ (Esser) and to guide actions (Lindenberg), how are these two combined? As noted by Luhmann (1984: 157), in interaction people start building expectations of each others’ expectations, on the basis of observed actions. According to the notion of relational signaling (Lindenberg, 2000, 2003; Wittek 1999; Six, 2004) the actions that a trustee undertakes, triggered by a mental frame, in deliberation or automatic response, constitute relational signals that are observed and interpreted by the trustor.

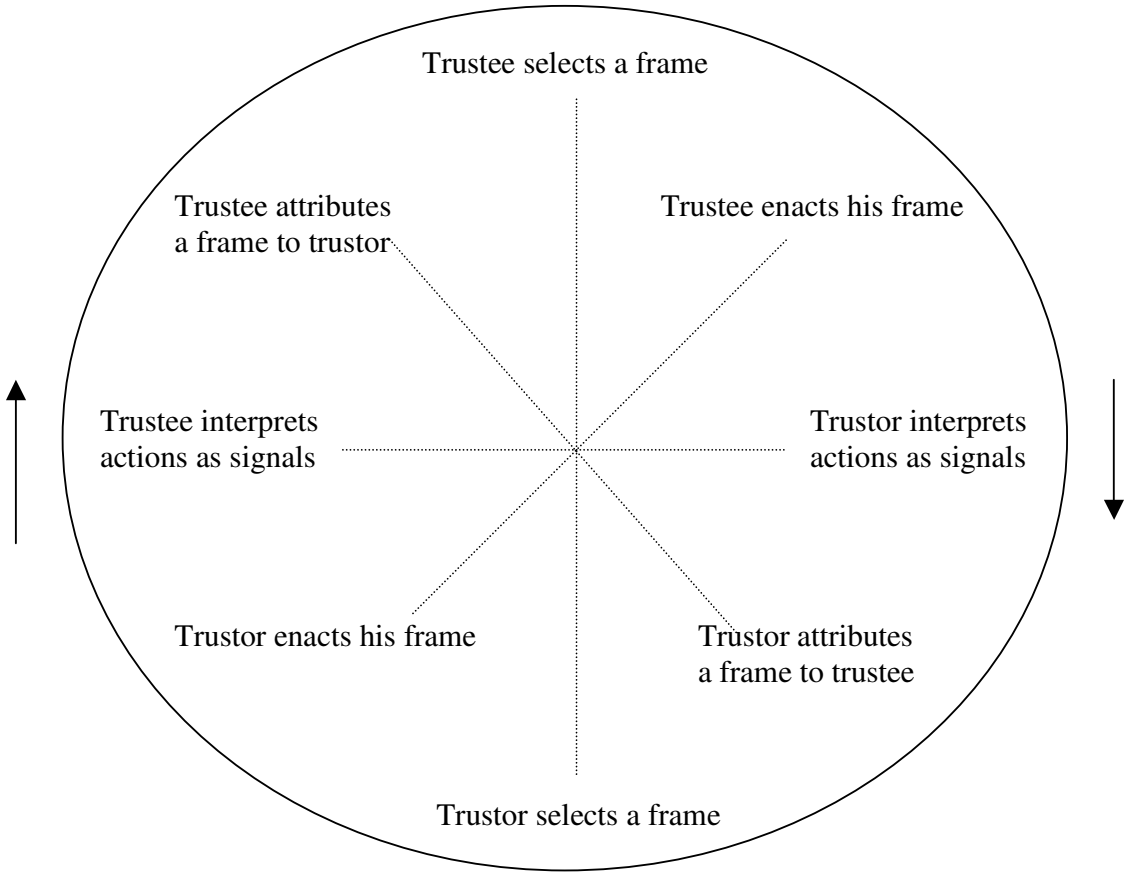
For frame selection I propose the following. The trustee selects a frame, which generates actions that function as signals to the trustor, who on the basis of these signals attributes a salient frame to the trustee and selects a frame for his own response, in the selection of a script, which generates actions taken as signals by the trustee, who attributes a frame to the trustor, and selects his own frame. This yields a cycle of selection and attribution, in ongoing interaction, as illustrated in Figure 1. Note that while a trustor (trustee) may select the same frame as the one attributed to the trustee (trustor), in what amounts to a ‘tit-for tat response’, this is not necessarily the case. One may persevere in acting benevolently in the face of opportunism, and one may opportunistically exploit the benevolent. Along this cycle, in deliberative response people may try to anticipate effects of actions, their signaling and the response in attribution, selection and action. This models Luhmann’s notion of the formation of expectations of expectations.

The following questions remain:

1. How, more precisely, do frame selection and attribution take place
2. How does frame selection lead to action?
3. What determines automatic or deliberative response (in selection and attribution)

Here, these questions cannot all be answered. For answers, use can be made of decision heuristics recognized in social psychology. For a survey see e.g. Bazerman (1998), and for further elaboration e.g. Kahneman, Slovic & Tversky (1982). Here, I reflect a little further on how frame selection and attribution might be modeled.

Figure 1 Cycle of frame selection and attribution

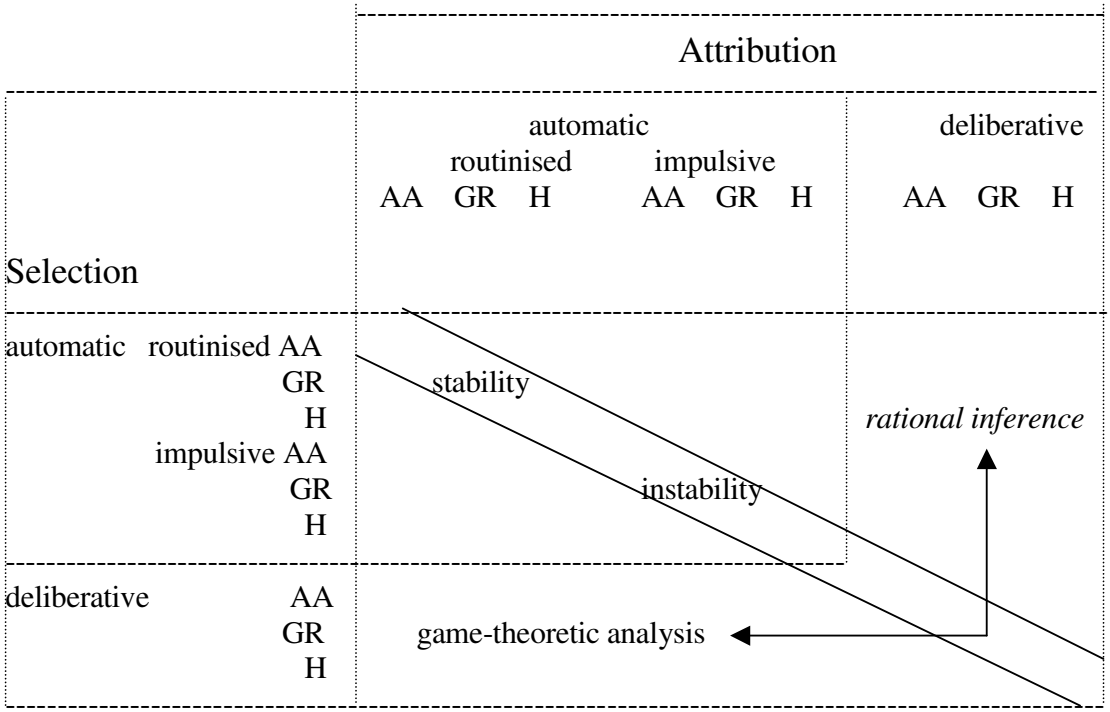


Selection and attribution

The salience, and hence stability, of a frame, and the likelihood of switching to a subsidiary frame, depends on whether it is supported by those other frames. For example, acting appropriately, in a trustworthy fashion, is most stable when it also builds resources and satisfies hedonic drives. One will switch to a frame of self-interest when temptation or pressure exceeds one's ability to resist. Conversely, one will switch from a self-interested to an other-directed frame when threat or temptation subsides and loyalty assumes more prominence. Decision heuristics from social psychology may be used to understand how this happens (Nooteboom, 2002).
Attribution of a self-interested frame (H, GR) to the trustee seems likely to trigger the defensive selection of a similar frame by the trustor, particularly when the attribution is based on strong triggers ('availability') of fear of loss, in what amounts to a 'tit for tat' strategy. However, that is not necessarily the case, even when the attribution is automatic rather than deliberative. People may control a shock of fear of loss and stick to an other-directed frame (AA), in several ways. Firstly, such a response may be deliberative, in the realization that a misinterpretation may be at play, with a mis-attribution of opportunism where in fact a mishap or lack of competence may be the cause of failure. However, this may be a psychologically difficult feat to achieve, and one may need the sobering caution

from a third party or go-between. See Nooteboom (2002) for an analysis of roles that go-betweens can play in the building and maintenance of trust.

Table 1 Attribution and selection



The trustor may respond with a different frame from the one he attributed to the trustee, and both attribution and selection may be automatic, in the two ways of routinised or impulsive response, or deliberative. Three frames for attribution and selection (AA, GR, H), in three modes (routinised, impulsive, deliberative) yield 81 logically possible action-response combinations, as illustrated in Table 1.

Deliberative attribution entails rational inference of scripts and corresponding frames, and deliberative selection typically entails game-theoretic type analysis of projected response in attribution to chosen actions. Here, the connection between action scripts and mental frames may be confounded in 'interest seeking with guile': one may choose actions that belong to scripts that enact an AA frame, while in fact one's salient frame is GR.

Impulsive attribution combined with impulsive frame selection will tend to yield instable relations, while routinised attribution in combination with routinised selection, if attributed and selected frames are the same (lie on the diagonal of the table) is likely to result more in stable relations.

The analysis demonstrates the importance of empathy, for correct attribution, on the basis of knowledge of the trustee's idiosyncracies of conduct and thought, and his strengths and weaknesses, in competence, loyalty, and resistance to temptation and pressures of survival.

For example, one may try to interpret an action as enacting the frame of acting appropriately. For example, the trustee's openness about a mistake is seen as fitting into the set of actions that belong to acting appropriately. In deliberate attribution one carefully tests assumptions concerning the attribution of a frame, considering whether other actions confirm that frame, and whether the action may also fit alternative frames. In routine attribution one attributes without much consideration, according to past anchors, and in impulsive attribution one tries to fit actions into frames that surge to attention as 'available' on the basis of fear or other emotion.
From interaction, including the disappointment of expectations, one may learn and innovate in several ways. One may discover new variations upon existing repertoires of actions associated with a frame, a new allocation of actions across mental frames, novel actions or even novel mental frames. This learning may serve for a better attribution of frames to trustees, and for an extension of one's own repertoires of action and mental frames. Here, even the breach of trust may be positive, as a learning experience, and may be experienced as such.

Further research

I have only been able to give a rough sketch of how human nature, as explained in social psychology, may provide a basis for modeling social-cognitive processes in agent-based models in general, and in the build-up and adaptation of trust in particular. Much work remains to be done in translating this into model design.
In particular, we need to fill in the details of how frame attribution and selection take place, in Figure 1 and Table 1. This may be based, in more detail, on decision heuristics identified in social psychology.
However, as recognized also by Marsella et al. (2004), there is the usual trade-off to be considered between detail and management of complexity. While, as Marsella et al. say, complexity may lie in the detail with which agents are modeled, this is feasible and desirable only with very few interacting agents, while in other studies complexity is emergent from the system of interaction between many agents, more simply modeled.

References

Arthur, W. Brian (1991). Designing economic agents that act like human agents: A behavioural approach to bounded rationality. American Economic Review, 81/2, 353 -359.

Arthur, W. Brian (1992). Designing economic agents that behave like human agents. Journal of Evolutionary Economics, 3/1, 1-22.

Bazerman, M. (1998). Judgement in managerial decision making, New York: Wiley.

Breban, S. (2002). A coalition formation mechanism based on inter-agent trust relationships, AAMAS conference, July 15-19, Bologna.

Brichoux, D. & P.E. Johnson (2002). The power of commitment in cooperative social action, Journal of Artificial Societies and Social Simulation, 5/3. http://jass.soc.surrey.ac.uk/5/3/1.html

Castelfranchi, C. & R. Falcone (1999). Social trust: A cognitive approach, CAiSE conference, Heidelberg.

Diekmann A. & W. Przepiorka (2005). The evolution of trust and reputation: Results from simulation experiments, unpublished paper, Dept. of Humanities, Social and Political Sciences, Swiss Federal Institute of Technology, Zürich.

Esser, H. (2005). 'Rationalität und Bindung – Das Modell der Frame-Selektion und die Erklärung des normativen Handlens', in M. Held, G. Kubon-Gilke, R. Sturm (eds), Normative und institutionelle Grundfragen der Okonomik, Jahrbuch 4, Reputation und Vertrauen, Marburg: Metropolis, 85-112.

Gambetta, D. (1988). Can we trust trust?, in: D. Gambetta (ed.): Trust; making and breaking of cooperative relations, Oxford: Blackwell, 213-237.

Gans, G., M. Jarke, S. Kethers, G. Lakemeyer, L. Ellrich, C. Funken & M. Meister (2001). Towards (dis)trust-based simulations of agent networks, proceedings of the 4th workshop on deception, fraud and trust in agent societies, Montreal, 13-25.

Gorobets, A. & B. Nooteboom (2005). Adaptive build-up and break-down of trust: An agent based computational approach, paper in review.

Gulati, R. (1995). Does familiarity breed trust? The implications of repeated ties for contractual choice in alliances, Academy of Management Journal, 30/1, 85-112.

Khaneman, D., P. Slovic & A. Tversky (eds 1982). Judgment under uncertainty: Heuristics and biases, Cambridge UK: Cambridge University Press.

Kirman, A.P. & N.J. Vriend (2001). Evolving market structure: An ACE model of price dispersion and loyalty. Journal of Economic Dynamics and Control ,25/3&4, 459-502.

Klos, T.B. & B. Nooteboom (2001). Agent-based computational transaction cost economics, Journal of Economic Dynamics and Control, 25, 503-526.

Lane, David A. (1993). Artificial worlds and economics, part II. Journal of Evolutionary Economics, 3/3, 177-197.

Lewicki, R.J. & B.B. Bunker (1996). Developing and maintaining trust in work relationships, in R.M. Kramer and T.R. Tyler (eds), Trust in organizations: Frontiers of theory and research, Thousand Oaks: Sage, 114-139.

Lindenberg, S. (2000). It takes both trust and lack of mistrust: The workings of cooperation and relational signalling in contractual relationships, Journal of Management and Governance, 4, 11-33.

Lindenberg, S. (2003). Governance seen from a framing point of view: the employment relationship and relational signalling. In: B. Nooteboom and F.E. Six, 2003. The trust process; Empirical studies of the determinants and the process of trust development. Cheltenham UK: Edward Elgar, 37-57.

Luhmann, N. (1980). Rechtssociologie, 2, extended edition, Reinbeck bei Hamburg.

Marsella, S.C., D.V. Pynadath & S.J. Read (2004). PsychSim: Agent-based modeling of social interactions and influence, proceedings International Conference on Cognitive Modelling, Mahwah, Earlbaum, 243-248.

Maximilien, E.M. & M.P. Singh. (2005). Agent-based trust model involving multiple qualities, AAMAS conference, July 25-29, Utrecht, Netherlands.

McAllister, D.J. (1995). Affect- and cognition based trust as foundations for interpersonal cooperation in organizations, Academy of Management Journal, 38/1, 24-59.

Meijer, S. & T. Vervaart (2005). Feasibility of multi-agent simulation for the trust and tracing game, in M. Ali & F. Esposito (eds.), Proceedings IEA/AIE, Heidelberg: Springer, 145-154.

Minsky, M. (1975). A framework for representing knowledge, in P.H. Winston (ed.), The psychology of computer vision, NY: McGraw-Hill.

Muller, P (2003). On reputation, leadership and communities of practice, EAEPE conference, 7-9 November, Maastricht

Nooteboom, B. (1999). Inter-firm alliances: Analysis and design, London: Routledge.

Nooteboom, B. (2000). Learning and innovation in organizations and economies, Oxford: Oxford University Press.

Nooteboom, B. (2002). Trust: Forms, functions, foundations, failures and figures, Cheltenham UK: Edward Elgar.

Nooteboom, B. (2004). Inter-firm collaboration, learning and networks: An integrated approach, London: Routledge.

Pahl-Wostl, C. & E. Ebenhöh (2004). Heuristics to characterise human behaviour in agent based models, in C. Pahl-Wostl, S. Schmidt, A.E. Rizzoli & A.J. Jakeman (eds.), Complexity and integrated resources management, proceedings 2nd biennial conference IEMSS, 14-17 June, Osnabrück, 177-184.

Pettit, Ph (1995). The virtual reality of homo economicus, The Monist, 78/3, 308-329.

Reich, W. (2004). Reasoning about other agents: A plea for logic-based methods, Journal of Artificial Societies and Social Simulation, 7/4. http://jass.soc.surrey.ac.uk/7/4/4.html

Six, F. (2004). Trust and trouble; Building interpersonal trust within organizations, PhD dissertation, Erasmus University Rotterdam.

Smith Ring, P. & A. van de Ven (1994). Developmental processes of cooperative interorganizational relationships, Academy of Management Review, 19/1, 90 - 118.

Tversky, A. & Kahneman, D. (1983). Probability, representativeness, and the conjunction fallacy, Psychological Review, 90/4, 293-315.

Williamson, O.E. (1975). Markets and hierarchies, New York: Free Press.

Williamson, O.E. (1993). Calculativeness, trust, and economic organization, Journal of Law & Economics 36, 453-486.

Wittek, R.P.M. (1999). Interdependence and informal control in organizations, PhD dissertation. University of Groningen, the Netherlands.

Younger, S. (2005). Reciprocity, sanctions, and the development of mutual obligation in egalitarian societies, Journal of Artificial Societies and Social Simulation, http://jass.soc.surrey.ac.uk/8/2/9.html

Zand, D.E. (1972). Trust and managerial problem solving, Administrative Science Quarterly, 17/2: 229 - 239.

Zucker, L.G. (1986). 'Production of trust: Institutional sources of economic structure', in Barry, Staw & Cummings, Research in organisational behaviour, 8, 53-111.

Appendix A: Details of the Model

Preference and matching

Preference is specified as follows:

$$\text{score}_{ij} = \text{profitability}_{ij}^{\alpha_i} \cdot \text{trust}_{ij}^{1-\alpha_i} \qquad (1)$$

where: score_{ij} is the score i assigns to j, $\text{profitability}_{ij}$ is the profit i can potentially make 'through' j, trust_{ij} is i's trust in j and $\alpha_i \in [0, 1]$ is the weight i attaches to profitability relative to trust, i.e. the 'profit-elasticity' of the scores that i assigns; i may adapt the value of α_i from each timestep to the next.
At each time step, all buyers and suppliers establish a strict preference ranking over all their alternatives. Random draws are used to settle the ranking of alternatives with equal scores. The matching of partners is modeled as follows. On the basis of preferences buyers are assigned to suppliers or to themselves, respectively. When a buyer is assigned to himself this means that he makes rather than buys. In addition to a preference ranking, each agent has a 'minimum tolerance level' that determines which partners are acceptable. Each agent also has a quota for a maximum number of matches it can be involved in at any one time. A buyer's minimum acceptance level of suppliers is the score that the buyer would attach to himself. Since it is reasonable that he completely trusts himself, trust is set at its maximum of 1, and the role of trust in the score is ignored: $\alpha = 1$. The algorithm used for matching is a modification of Tesfatsion's (1997) deferred choice and refusal (DCR) algorithm and it proceeds in a finite number of steps, as follows:

1. Each buyer sends a maximum of o_i requests to its most preferred, acceptable suppliers.
2. Each supplier 'provisionally accepts' a maximum of a_j requests from its most preferred buyers and rejects the rest (if any).
3. Each buyer that was rejected in any step fills its quota o_i in the next step by sending requests to next most preferred, acceptable suppliers that it has not yet sent a request to.
4. Each supplier again provisionally accepts the requests from up to a maximum of a_j most preferred buyers from among newly received and previously provisionally accepted requests and rejects the rest. As long as one or more buyers have been rejected, the algorithm goes back to step 3.

The algorithm stops if no buyer sends a request that is rejected. All provisionally accepted requests are then definitely accepted.

Trust and trustworthiness

An agent i's trust in another agent j depends on what that trust was at the start of their current relation and on the past duration of their current relation:

$$t_i^j = t_{\mathrm{init},i}^j + (1 - t_{\mathrm{init},i}^j)\left(1 - \frac{1}{fx + 1 - f}\right), \qquad (2)$$

where t_i^j = agent *i*'s trust in agent *j,*

$t_{\mathrm{init},i}^j$ = agent *i*'s initial trust in agent *j,*

x = the past duration of the current relation between agents *i* and *j, and*

f = trustFactor.

This function is taken simply because it yields a curve that increases with decreasing returns, as a function of duration x, with 100% trust as the limit, and the speed of increase determined by the parameter f .

In addition, there is a base level of trust, which reflects an institutional feature of a society. If an agent *j*, involved in a relation with an agent *i,* breaks their relation, then this is interpreted as opportunistic behavior and *i'*s trust in *j* decreases; in effect, *i*'s trust drops by a percentage of the distance between the current level and the base level of trust; it stays there as *i*'s new initial trust in *j,* $t_{\mathrm{init},i}^j$ until the next time *i* and *j* are matched, after which is starts to increase again for as long as the relation lasts without interruption.

The other side of the coin is, of course, one's own trustworthiness. This is modelled as a threshold τ for defection. One defects only if the advantage over one's current partner exceeds that threshold. It reflects that trustworthiness has its limits, and that trust should recognize this and not become blind (Pettit 1995, Nooteboom 2002). The threshold is adaptive, as a function of realized profit.

Costs and profits

Buyers may increase gross profits by selling more differentiated products, and suppliers may reduce costs by generating production efficiencies. There are two sources of production efficiency: economy of scale from a supplier producing for multiple buyers, and learning by cooperation in ongoing production relations. Economy of scale can be reaped only in production with general-purpose assets, and learning by cooperation only in production that is specific for a given buyer, with buyer-specific assets.

We assume a connection between the differentiation of a buyer's product and the specificity of the assets required to produce it. In fact, we assume that the percentage of specific products is equal to the percentage of dedicated assets. This is expressed in a variable $d_i \in [0, 1]$. It determines both the profit the buyer will make when selling his products and the degree to which assets are specific, which determines opportunities for economy of scale and learning by cooperation.

Economy of scale is achieved when a supplier produces for multiple buyers. To the extent that assets are specific, for differentiated products, they cannot be used for production for other buyers. To the extent that products are general purpose, i.e. production is not differentiated, assets can be switched to produce for other buyers. In sum, economy of scale, in production for multiple buyers, can only be achieved for the non-differentiated, non-specific part of production, and economy by learning by cooperation can only be achieved for the other, specific part.

Both the scale and learning effects are modelled as follows:

$$y = \max\left(0{,}1 - \frac{1}{fx + 1 - f}\right), \qquad (3)$$

where:
for the scale effect, *f*=scaleFactor, *x* is general-purpose assets of supplier *j* summed over all his buyers and scale efficiency $y = e_{s,j}$;

for the learning effect, *f*=learnFactor; *x* is the number of consecutive matches between supplier *j* and buyer *i* and learning efficiency $y = e^{i}_{l,j}$.

Formula (3) expresses decreasing returns for both scale and experience effects. For the scale effect, it shows positive values along the vertical axis only for more than 1 general-purpose asset. This specifies that a supplier can be more scale-efficient than a buyer producing for himself only if the scale at which he produces is larger than the maximum scale at which a buyer might produce for himself. For the learning effect, a supplier's buyer-specific efficiency is 0 in their first transaction, and only starts to increase if the number of transactions is larger than 1. If a relation breaks, the supplier's efficiency due to his experience with the buyer drops to zero. The resulting specification of profit is specified as follows:

Adaptation

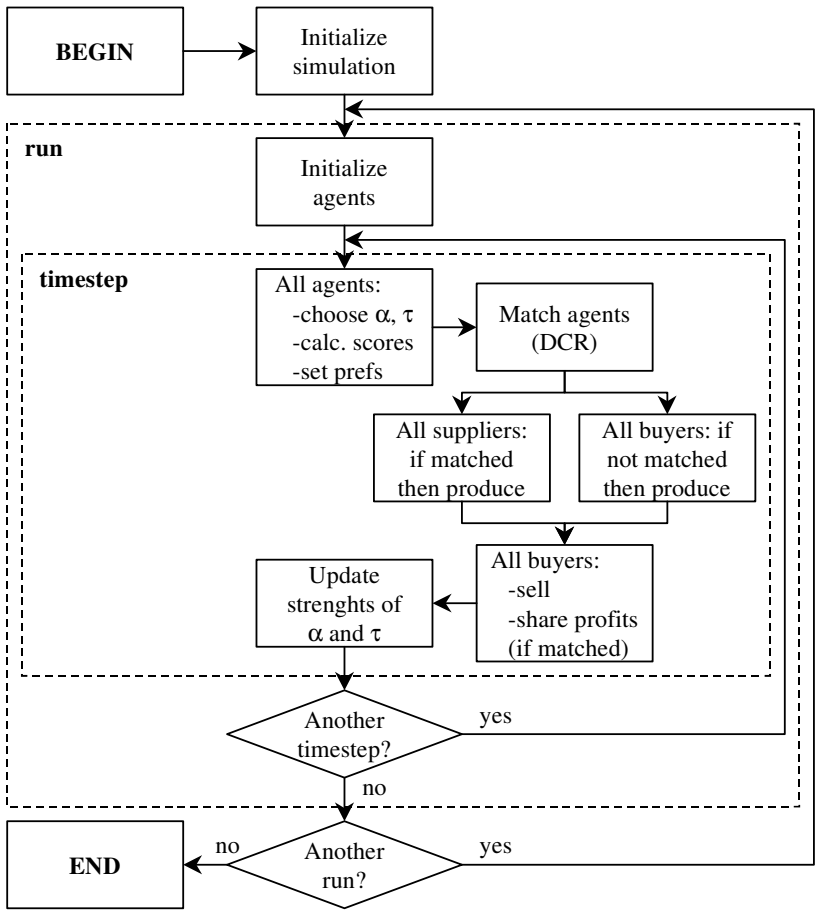
Agents adapt the values for $\alpha \in$ [0, 1] (weight attached to profit relative to trust) and τ [0, 0.5] (threshold of defection) from one time step to the next, which may lead to changes in the scores they assign to different agents. Here, adaptation takes place on the basis of past, realized profit. While τ could conceivably rise up to 1, a maximum of 0.5 was set because initial simulations showed that otherwise relations would get locked into initial situations, with little switching. Note that this biases the model in favour of opportunism. At each time step, each agent assigns a 'strength' to each possible value of α and τ. This expresses the agent's confidence in the success of using that particular value. The various strengths always add up to constants C_α and C_τ, respectively. At the start of each timestep, the selection of values for α and τ is stochastic, with selection probabilities equal to relative strengths, i.e. strengths divided by C_α and C_τ, respectively. The strengths of the values that were chosen for α and τ at the start of a particular timestep are updated at the end of that timestep, on the basis of the agent's performance during that timestep, in terms of realized profit: the agent adds the profit obtained during the timestep to the strengths of the values that were used for α or τ. After this, all strengths are renormalized to sum to C_α and C_τ again (Arthur 1993). The idea is that the strength of values that have led to high performance (profit) increases, yielding a higher probability that those values will be selected again. This is a simple model of 'reinforcement learning' (Arthur 1991, Arthur 1993, Kirman and Vriend 2000, Lane 1993).

The algorithm

The algorithm of the simulation is presented by the flowchart in Fig.1. This figure shows how the main loop is executed in a sequence of discrete time steps, called a 'run'. Each

simulation may be repeated several times as multiple runs, to even out the influence of random draws in the adaptation process. At the beginning of a simulation, starting values are set for certain model parameters. The user is prompted to supply the number of buyers and suppliers, as well as the number of runs, and the number of timesteps in each run. At the start of each run, all agents are initialized, e.g. with starting values for trust, and selection probabilities for α and τ. In each timestep, before the matching, each agent chooses values for α and τ, calculates scores and sets preferences. Then the matching algorithm is applied. In the matching, agents may start a relation, continue a relation and break a relation. A relation is broken if, during the matching, a buyer does not send any more requests to the supplier, or he does, but the supplier rejects them.

Fig. A1. Flowchart of the simulation.



After matching, suppliers that are matched to buyers produce and deliver for their buyers, while suppliers that are not matched do nothing. Buyers that are not matched to suppliers produce for themselves ('self-matched', in 'make' rather than 'buy'). Afterward, all buyers sell their products on the final-goods market. Profit is shared equally with their supplier, if they have one. Finally, all agents use that profit to update their preference rankings (via α and τ), used as input for the matching algorithm in the next timestep.

Across timesteps realized profits are accumulated for all buyers and suppliers, and all the relevant parameters are tracked.
Note that, by implication, suppliers may fail to produce, and then have zero profit. Thus, there is no explicit mechanism of death. However, the procedure may be interpreted as exit of all suppliers with zero profit, accompanied by potential entry on new suppliers, announcing their readiness to give quotes to buyers, up to the maximum number of suppliers specified for the run. Note also that it is conceivable, given the logic of matching, that a supplier breaks with a buyer in his aim to go for a more attractive one, then lose the bidding for that buyer, and be left empty-handed. Then, it would be more reasonable for the supplier to first verify his goal attainment before breaking his existing relationship. However, in a large set of simulations, across a wide area of parameter space, this happened only once, at a very high level of opportunism, and it may not be unrealistic that sometimes such error is made, in an over-eagerness to switch to a more attractive partner.

Appendix B: Details of Simulation Outcomes

High initial trust dictates buy relative to make for all levels of specific investments. For high specific investments, buyers' maximum profit is almost the same as in the cases of average or low initial trust. Low initial trust imposes make relative to buy, but buyers' maximum profits for low specific investments are smaller than in the case of high initial trust. Overall, across all parameter settings, profit tends to be higher under high than under low trust.
Under medium or low trust, high product differentiation favours make relative to buy because the switching cost is larger and there is less potential for economy of scale. But if learning by cooperation becomes stronger, relative to scale effects, buyers employ that advantage in a strategy of ongoing relations with suppliers, and achieve a higher profit than when they make themselves. If agents put their emphasis on trust (α=0) and loyalty (τ=0.5) buyers get a big advantage in the terms of profit for high specific investments by following the strategy of learning by cooperation. If agents focus on profitability rather than on trust (α=1) and neglect loyalty (opportunistic, τ=0) buyers get some advantage for low specific investments by following the scale strategy (50%) and producing themselves (50%). But if agents are loyalists (τ=0.5) buyers get an advantage for both low and average d by following the scale strategy (60% and 40% respectively) and producing themselves (40% and 60% respectively). Generally, under low trust and low weight attached to trust, buyers forego opportunities for collaboration that may yield learning by cooperation. In sum, high initial trust favours outsourcing ('buy') and it gives an advantage for all agents in comparison with low initial trust, where buyers get a smaller profit by insourcing ('make').
In addition to the expected results, the model yields a few unanticipated results. One is that buyers organize closer to maximum possible efficiency for high levels of specific investments/specialization. The reason is that for low levels of specific investments there is more scope for scale effects, but this is difficult to attain by having suppliers supply to the maximum number of buyers. A strong effect of learning by cooperation, a high weight attached to trust, and high loyalty favour the learning by cooperation strategy for

high levels of specific investments, while a high weight attached to profit and high loyalty favour the scale strategy for low and average levels of specific investments. Finally, it is not always the case that a high weight attached to profitability relative to trust favours opportunism. Once a buyer begins to profit from learning by cooperation, an emphasis on profit may also lead to loyalty, in an ongoing relationship.