Oxford Handbooks Online

Trust Games: Game-Theoretic Approaches to Embedded Trust

Vincent Buskens, Vincenz Frey, and Werner Raub

The Oxford Handbook of Social and Political Trust
Edited by Eric M. Uslaner

Print Publication Date: Mar 2018
Subject: Political Science, Political Behavior, Political Methodology
Online Publication Date: Jan 2017 DOI: 10.1093/oxfordhb/9780190274801.013.38

Abstract and Keywords

This article offers an overview of different variants of trust games and shows how game-theoretic modeling can contribute to an analysis of conditions for placing and honoring trust in such games. The focus is on explaining trust rather than on explaining consequences of trust for individual behavior or for outcomes such as societal cohesion or economic prosperity. Specifically, game-theoretic modeling allows for analyzing how the “embeddedness” of trust games in long-term relations between actors and in networks of relations can be a basis for informal norms and institutions of trust. Game-theoretic modeling also allows for analyzing actors’ incentives to modify embeddedness characteristics so that informal norms and institutions of trust become feasible. We discuss how game-theoretic models can be used to derive testable predictions for experiments with trust games and sketch empirical evidence from such experiments.

Keywords: Trust, trust games, game theory, embeddedness, norms, institutions, experiments, reputation

Introduction

CONSIDER an example of social exchange: Ego helps Alter today, assuming that Alter will help Ego tomorrow. If Alter indeed provides help tomorrow, both Ego and Alter are better off than they would be without helping each other. However, Alter may be tempted to benefit from Ego’s help today without providing help himself tomorrow. Ego may anticipate this and not provide help in the first place, leaving both Ego and Alter worse off than if they had helped each other.

In economic exchange—transactions in online markets being an example—a buyer may be insufficiently informed about the quality of a good offered by a seller. If this buyer decides to buy and the seller delivers a good of adequate quality, both buyer and seller are better off than without a transaction. However, the seller may be tempted to secure an extra profit by selling a bad product for the price of a good one, leaving the buyer worse off than if she had decided not to buy. Anticipating this, the buyer may abstain from entering the transaction, again leaving buyer and seller worse off than after a "smooth" transaction.

A similar problem can arise if a voter decides on casting a vote for a representative who might later choose to initiate or support policies other than those preferred by the voter. The representative can benefit from changing his position opportunistically and the voter may regret having voted for this representative.

An experimental design modeling core features of these examples involves two subjects A and B. A can decide to transfer an endowment of, say, 20 points (converted into money at a fixed exchange rate at the end of the experiment) to subject B. If A decides to transfer, the experimenter triples the endowment. Subject B then chooses between returning 30 points to A or keeping 60 points for himself. If B splits, both A and B are **(p. 306)** better off than had A decided not to transfer her endowment. However, keeping 60 points for himself ensures an even higher payoff from the experiment for B than splitting.

The term "trust," as this *Handbook* shows, can have different meanings. Trust games are workhorses to study trust in the sense of our examples. Game theory provides a set of theoretical tools—concepts and assumptions—to model and analyze situations like those highlighted in our examples. Behavioral game theory "expands analytical theory by adding emotion, mistakes, limited foresight, doubts about how smart others are, and learning to analytical game theory . . .. Behavioral game theory is . . . an approach . . . which uses psychological regularity to suggest ways to weaken rationality assumptions and extend theory" (Camerer 2003, 3). Much work in behavioral game theory involves research on trust (see, e.g., Camerer 2003, chap. 2.7). Experiments are widely used to systematically test empirical predictions derived from game-theoretic and other models, including predictions on behavior in trust situations. Experiments likewise yield insights on empirical regularities that subsequent theory development can try to explain.

Research on trust focuses on two different issues (see Craswell 1993). First, in the spirit of Arrow's (1974, 24) well-known remark on trust as "an important lubricant of a social system," research addresses the *consequences of trust*, be it consequences on the level of individual behavior or consequences of trust for more macro-level outcomes such as societal cohesion or economic prosperity. Assumptions on trust are then used to help explain other phenomena. In such research, trust is part of the *explanans*. Conversely, one can study the *determinants of trust*: What are conditions that foster trust in social and economic exchange? Trust is the *explanandum* in this research. Trust games are used, at least primarily, to study trust as an explanandum by focusing on conditions that favor or undermine trust.

This article provides an overview of how trust games can be used to study trust as an explanandum. More specifically, we discuss how game-theoretic reasoning and experimental research can be employed for understanding how macro-conditions affect trust. For example, how do macro-conditions such as the stability over time of relations between actors affect their behavior in trust situations? How about effects of networks of relations between actors? How can such relations and networks of relations foster trust through self-enforcing norms and how can they give rise to informal institutions that foster trust? We also discuss conditions such that actors themselves, anticipating that trust can make them better off but may not be easily attained, modify their networks so that trust is facilitated.

The article continues with an overview of different variants of trust games. An overview follows of some theoretical models on how macro-conditions affect trust. Subsequently, we outline experimental evidence. Concluding remarks follow.

Variants of Trust Games

We follow Coleman's (1990, 96–99) general characterization of social and economic situations involving trust like those in our introduction. In such situations, two actors (p. 307) are involved, a trustor and a trustee. In our examples, Ego, the buyer, the voter, and subject A are trustors. Alter, the seller, the representative, and subject B are trustees. In each case, the trustor must decide whether or not she[1] places trust, that is, whether or not to help Alter, to buy, to vote for the representative, or to transfer the endowment.

Coleman (1990, 97–99) sketches four features of trust situations. First, placing trust implies that the trustee can subsequently honor or abuse trust. Alter can honor trust by helping Ego tomorrow, the seller by selling a good of adequate quality, the representative by being consistent with policies he favored in his election campaign, and subject B by returning half of the tripled endowment to A. Second, if the trustee honors trust, the trustor is better off than if trust were not placed. Conversely, the trustee can abuse trust. Alter can refuse to provide help tomorrow, the seller can sell a bad product for the price of a good one, the representative can deviate from what he promised in his election campaign, and subject B can keep the tripled endowment for himself. If trust is abused, the trustor is worse off than had trust not been placed. Third, through placing trust, the trustor transfers resources to the trustee without any "real commitment" (Coleman 1990, 98) of the trustee to honor trust. Fourth, there is a time lag between the actions of trustor and trustee. The trustor first decides on placing or not placing trust, while the trustee only acts in the future, so that the trustor cannot know for sure but has to anticipate whether or not the trustee would honor trust. A fifth feature, less explicitly addressed by Coleman, is that the trustee may have an incentive to abuse trust, at least in the short run, since he may benefit in financial or other terms from doing so. A sixth feature, likewise reflected in our examples, is that the trustor and trustee are both better off if trust is placed and honored than if the trustor does not place trust.

The simplest game-theoretic model for trust situations considered here is the standard *Trust Game* (TG) depicted in Figure 14.1 (Dasgupta 1988; Kreps 1990; a tree-like representation like in Figure 14.1 is known as the "extensive form" of a game). To facilitate interpretation, Figure 14.1 also includes a numerical example representing the experimental setup described above. The game involves a trustor (indexed "1") and a trustee (indexed "2"). The trustor moves first and has to choose between placing and not placing trust (transferring or not transferring 20 points). The interaction ends if trust is not placed. In this case, the trustor receives payoff P_1, while the trustee receives payoff P_2 (with P_1 equal to 20 and P_2 equal to 0 points in the example). If (p. 308) trust is placed, the trustee moves, choosing between honoring and abusing trust (sharing or not sharing the points). The interaction ends thereafter. Honored trust implies payoffs $R_i > P_i$, $i = 1, 2$. Abused trust is associated with payoffs $S_1 < P_1$ for the trustor and $T_2 > R_2$ for the trustee. In the example in Figure 14.1, R_i is equal to 30 points, S_1 equal to 0 points, and T_2 equal to 60 points. Note that the TG captures core features of trust problems as in our examples. We refer to placing trust also as *trustfulness* and to honoring trust as *trustworthiness*, with "trust" sometimes referring to a situation resembling a trust game and sometimes referring to trustful and trustworthy behavior in such a situation.

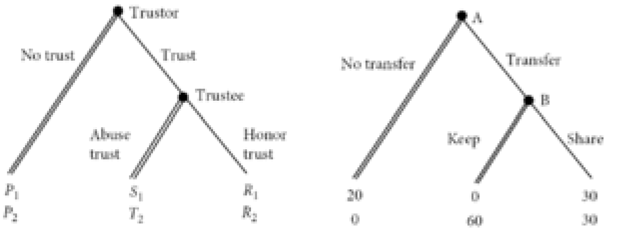

Figure 14.1 The standard Trust Game $S_1 < P_1 < R_1, P_2 < R_2 < T_2$ and an example.

Standard game-theoretic assumptions[2] include that the payoffs actors receive at the end of the game represent utilities (subsequently, we typically assume cardinal utilities). Also, all actors know the structure of the game (they know Figure 14.1) and also know that all actors know the structure of the game and so forth ("common knowledge"). The game is played noncooperatively in the sense that actors are unable to make enforceable agreements or to incur commitments that are not explicitly modeled as moves in the game (compare Coleman's point that there is no "real commitment" for the trustee). The actors are rational in the sense that they maximize utility, given their expectations on the behavior of other actors, and actors assume that other actors are rational as well. Under these assumptions and, since $T_2 > R_2$, the trustee would abuse trust if the trustor places trust. The trustor anticipates this and hence, since $P_1 > S_1$, does not place trust in the first place. In game-theoretic terms, not placing trust, while placed trust would be abused, is the unique subgame perfect equilibrium of the game; that is, for each situation that could emerge during the game, each actor's strategy maximizes that actor's payoff, given the strategy of the other actor (in the following, we employ this equilibrium concept and sometimes concepts that satisfy additional criteria). The equilibrium is indicated by double lines in Figure 14.1.

The TG reflects Coleman's (1990, 99) point that placing trust is risky for the trustor: she is better off when trust is placed and honored than when she does not place trust, but the trustor regrets having placed trust if the trustee abuses trust. Since $R_i > P_i$, the equilibrium outcome of the TG—no trust placed—is worse for both actors than the outcome such that trust is placed and honored, but the outcome with trust placed and honored is not a result of equilibrium behavior and is thus not attainable for rational actors. In terms of the example in Figure 14.1, subject A anticipates that B prefers 60 points over 30 and, thus, would leave A with 0 points if A would transfer. Therefore, A prefers not to transfer, resulting in 20 points for A and 0 for B. For both subjects, this is worse than the outcome with 30 points for each when A transfers and B shares. In this sense, a trust situation represents a problem for each of the actors. The TG is a social dilemma: individual rationality induces an outcome that is worse for each actor than a "collectively rational" outcome (Rapoport 1974). The problem inherent in trust situations in this article is due to the incentives of the trustee for "opportunistic behavior" in the sense of abusing trust. This has to be distinguished from situations such that the second actor may lack the abilities and competencies to realize a beneficial outcome (p. 309) for the first actor. In the literature (e.g., Barber 1983), the latter case has been discussed as a problem of *confidence*.

There are various other, more complex versions of trust games. Considering such extensions is useful because they allow accounting for additional features of trust problems that may be relevant for applications. Furthermore, such extensions allow for checking the robustness of predictions to variations in assumptions. We discuss two extensions. One extension includes information problems, namely, incomplete information of the trustor concerning characteristics of the trustee (Camerer and Weigelt 1988; Dasgupta 1988). Assume that the trustor does not know for sure about the incentives of the trustee. More precisely, the trustee could be one of two "types." For an *unreliable* trustee, it pays off to abuse trust, since $T_2 > R_2$ as in the TG. However, a *reliable* trustee has no incentive to abuse trust. Rather, the payoff for such a trustee after abused trust is $T_2^* < R_2$, for example, because he does not care exclusively for his own material outcome but also derives disutility from a bad conscience due to having abused trust. In social exchange, for instance, Alter could feel guilty if he does not reciprocate the favor of Ego. Such a feeling of guilt may affect Alter so that he prefers helping in return. The trustee knows his own type, but the trustor only knows that she interacts with probability π ($0 < \pi < 1$) with a reliable trustee, while the trustee is unreliable with probability $1 - \pi$. Figure 14.2 shows a *Trust Game with incomplete information* (TGI). In the TGI, there is an initial move of Nature that determines the type of trustee. The trustor only knows the probability π but cannot observe the outcome of Nature's move (in Figure 14.2, encircling the two nodes where the trustor has to make a move indicates that the trustor cannot distinguish between these nodes).

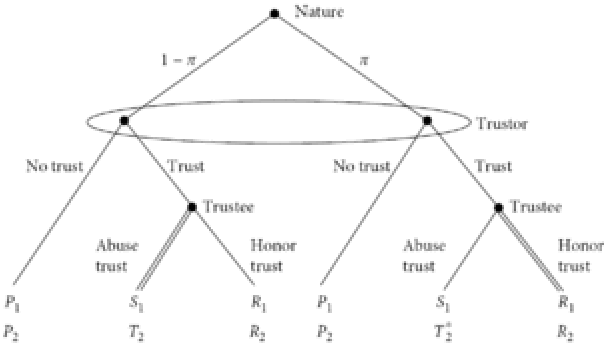

Figure 14.2 Trust Game with incomplete information $S_1 < P_1 < R_1, P_2 < R_2 < T_2,\ T_2^* < R_2$.

Equilibrium behavior in the TGI requires that an unreliable trustee would abuse trust (since $T_2 > R_2$), whereas a reliable trustee would honor trust (since $T_2^* < R$). For the trustor, not placing trust yields the certain payoff P_1, while the expected payoff (p. 310) associated with placing trust is $\pi R_1 + (1-\pi)S_1$. Hence, equilibrium behavior implies that the trustor places trust if $\pi R_1 + (1-\pi)S_1 > P_1$, which is equivalent with

$$\pi > (P_1 - S_1)/(R_1 - S_1)$$

(1)

The right-hand side of this inequality is a useful measure of the risk for the trustor and for the size of the trust problem.

The TGI represents Coleman’s (1990, 99) idea that the trustor has to decide about placing a bet. Coleman’s well-known condition for placing trust is in fact implied by condition (1). Hence, Coleman’s condition follows from a game-theoretic model.

Our second extension is the *Investment Game* (IG; Berg, Dickhaut, and McCabe 1995; Ortmann, Fitzgerald, and Boeing 2000). While the actors make binary choices in the TG and TGI, in the IG, the trustor chooses the degree to which she places trust and the trustee chooses the degree to which he honors trust. This is modeled by assuming that the trustor has an endowment E_1 and chooses an amount M_1 to send to the trustee ($0 \leq M_1 \leq E_1$). This “investment” M_1 is then multiplied by $m > 1$ and the trustee receives mM_1. The parameter m can be seen as indicating the returns from trade due to the trustor’s investment. Subsequently, the trustee chooses an amount K_2 he returns to the trustor, with $0 \leq K_2 \leq mM_1$. The game ends with the trustor receiving $V_1 = E_1 - M_1 + K_2$ and the trustee receiving $V_2 = mM_1 - K_2$. While M_1 indicates the trustor’s trustfulness, K_2 indicates how trustworthy the trustee is.[3]

In the IG, equilibrium behavior implies that the trustee would never return anything, that is, he would choose $K_2 = 0$ for all M_1, and that the trustor, anticipating this, sends nothing ($M_1 = 0$). The actors thus forgo all gains from trade: both trustor and trustee would be better off if the trustor would send $M_1 > 0$ and the trustee returns K_2 such that $mM_1 > K_2 > M_1$.[4]

Trust as Explanandum: Theory on How Social Conditions Affect Trustfulness and Trustworthiness

There is quite some theory on trust in “isolated encounters,” with trustor and trustee playing a focal trust game once and only once, without being able to condition behavior in future interactions on what happens in the trust game. Also, there are no previous interactions that have repercussions for the focal trust game. Such isolated encounters are typically studied in the laboratory, since it is hard to assure “isolation” using nonexperimental designs. Under standard game-theoretic assumptions sketched in the first section of this article, the strong prediction is that the trustor will not be trustful, while the trustee would not be trustworthy in the TG or the IG, nor in the TGI (p. 311) with $\pi < (P_1 - S_1)/(R_1 - S_1)$. Experiments show, however, that substantial percentages of subjects in the trustor role place trust and send positive amounts, while many subjects in the trustee role honor trust and return substantial amounts (for reviews, see, e.g., Camerer 2003, chap. 2.7; Johnson & Mislin 2011).

Models in the spirit of behavioral game theory can account for such empirical regularities by employing more complex assumptions than the standard ones. First, one could relax the rationality assumption. For example, from a bounded rationality perspective one could argue that subjects are used to repeated interactions in life outside the laboratory. As we will see, for example, in section 2.2, given repeated interactions, placing and honoring trust can be a result of equilibrium behavior. One then assumes that subjects follow rules of behavior in isolated encounters in an experiment that are appropriate when interactions are repeated (see Binmore 1998 for a discussion of such approaches).

Second, one could maintain the rationality assumption but modify the selfishness assumption (“utility = own money”). One would then assume that a subject’s utility associated with a certain outcome does not necessarily depend exclusively on, say, one’s own points received. Rather, subjects may have other-regarding preferences such that they care for the distribution of outcomes in addition to their own outcome (e.g., “inequity aversion”: a trustee’s utility may depend on his own payoff as well as on the difference between his own payoff and the trustor’s payoff). It is often argued (e.g., Fehr and Gintis 2007) that such preferences are the result of socialization processes and internalized norms and values. Roughly, if a trustee has indeed “suitable” other-regarding preferences and the trustor maintains beliefs that it is sufficiently likely that the trustee has such other-regarding preferences, placing and honoring trust can be the result of equilibrium behavior. In fact, a TGI is played in this case and trust is possible if π is large enough.

From a methodological perspective, assumptions on other-regarding preferences come with the risk that almost all behavior can be "explained" by assuming the "right" preferences and adjusting assumptions on utility accordingly. Therefore, assumptions on such preferences should be parsimonious and allow for explaining behavior in a broad range of different experimental games. Various models for other-regarding preferences are meanwhile available that do indeed account with the same set of assumptions for behavioral regularities, not only in trust games but also in other social dilemma games; in games involving distribution problems, such as the Ultimatum and Dictator Game; and in market games (for overviews, see Camerer 2003; Charness and Shmidov 2013; Cooper and Kagel 2015; Fehr and Schmidt 1999; and see Wilson's article in this *Handbook* for an overview of work on trust in isolated encounters that addresses how behavior in such encounters is related to various macro-conditions, including conditions that might affect other-regarding preferences, and of nonexperimental work in this field).

In the following, we neglect trust in isolated encounters. Rather, we show how other extensions of standard game-theoretic models can shed light on behavior in **(p. 312)** trust games. We take into account that trust situations often occur in settings in which actors interact repeatedly or are connected through third parties. For example, Ego and Alter who can help each other are classmates, so that helping each other is repeatedly an issue. A buyer might know other customers of a seller from whom she is about to buy a product and can exchange information about the seller with these other customers. We do this by embedding trust games in settings such that behavior in a focal trust game can have repercussions for future interactions or that previous interactions affect the focal trust game. This allows addressing systematically how trustfulness and trustworthiness in trust games depend on certain macro-conditions.

Deriving testable hypotheses on trust from game-theoretic models

Coleman's (1990, chap. 1; see also Raub, Buskens, and Van Assen 2011) diagram for relating macro- and micro-level propositions in social science explanations, depicted in Figure 14.3, is useful for making the logic of game-theoretic explanations of trust, including trust in embedded settings, explicit and for making explicit how testable implications can be derived from game-theoretic models. Coleman's diagram distinguishes between macro- and micro-level in the sense that "macro" refers to properties of a social system, while "micro" refers to individuals. Macro-level properties need not be properties of "large" systems; they also include properties of "small" systems such as a dyad or a triad.

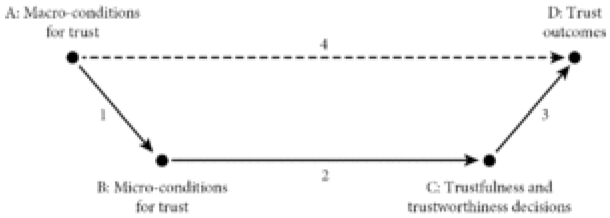
A: Macro-conditions for trust
D: Trust outcomes
4
1
3
2
B: Micro-conditions for trust
C: Trustfulness and trustworthiness decisions

Figure 14.3 Coleman's micro-macro diagram.

First, the specification of a game like our variants of trust games includes macro-conditions in the sense of opportunities and restrictions. These assumptions are represented by Node A in Coleman's diagram. They include, for example, the assumptions that the game involves two actors, that it is played noncooperatively, and assumptions on the sequence in which the actors move. The specification of the game also comprises assumptions on the actors' payoffs. These are micro-level assumptions related to Node B. Furthermore, the specification includes assumptions on macro-micro transitions (p. 313) that are summarized by the vertical Arrow 1 in the diagram. The specification shows how an actor's payoff depends on the behavior of the other actor and vice versa, that is, how the actors are interdependent. Next, rationality assumptions like the assumption of equilibrium behavior are represented by Arrow 2. Game-theoretic analysis then comprises deriving propositions on equilibria of the game and on properties of these equilibria. A simple example is that the TGI has an equilibrium such that trust is placed if condition (1) is fulfilled. This allows deriving implications concerning the behavior of rational actors, in our case, implications on trustfulness and trustworthiness. These implications are represented by Node C. Finally, one can derive propositions on macro-level effects such as that the outcome of the game leaves both actors worse off than another outcome they could have attained or, rather, that they manage to attain such a more beneficial outcome. A typical way of generating testable predictions is then to employ comparative statics analysis on equilibrium conditions such as condition (1), reasoning that behavior implied by a certain equilibrium becomes more (less) likely when the conditions for such an equilibrium become less (more) restrictive. For example, one would predict that trustfulness in the TGI becomes more likely when the trustor's risk as indicated by the right-hand side of (1) decreases or when the probability π of interacting with a reliable trustee increases. Condition (1) on the relation between the probability π and payoffs can be seen as a macro-condition characterizing how trustor and trustee are interdependent. We have thus derived predictions on how macro-conditions affect trustfulness.

Dyadic embeddedness

A focal trust game may be embedded (see Granovetter 1985 on the concept of "embeddedness"), in the sense that trustor and trustee interact repeatedly. Macro-conditions such as being related by friendship or family ties induce that Ego and Alter will contemplate repeatedly the exchange of social support. The same voters can vote again when a representative is up for reelection. To capture the effects of dyadic embeddedness in the sense of long-term relations between a trustor and trustee, consider *repeated trust games*. Two mechanisms can be distinguished through which dyadic embeddedness may affect behavior in a focal trust game: control and learning (Buskens and Raub 2002). *Control* refers to possibilities for the trustor to sanction the trustee in the future. The trustor can reward the trustee by again placing trust if trust would be honored in the focal trust game, and she can punish the trustee by not placing trust after trust would be abused. Information on behavior of the trustee from previous trust games can allow for *learning* of the trustor about unknown characteristics of the trustee and for updating her beliefs about the trustee.

A simple game-theoretic model of control effects through dyadic embeddedness is due to Kreps (1990). In this model, the TG from Figure 14.1 is played repeatedly in rounds 1, 2, . . ., t, ... After each round t the next round $t + 1$ is played with continuation probability w ($0 < w < 1$), while the repeated game ends after each round with (p. 314) probability $1 - w$. The focal TG is thus embedded in a more complex game in which the TG is repeated indefinitely often. In each round, actors can observe each other's behavior. An actor's expected payoff for the repeated game is the discounted sum of the actor's payoffs in each round, with the continuation probability w as discount parameter. For example, when trustor and trustee use strategies such that trust is placed and honored in each round, their expected payoff is $R_i + wR_i + \ldots + w^{t-1}R_i + \ldots = R_i/(1–w)$. Therefore, the larger the continuation probability w, the more an actor's payoff from the repeated game depends on what the actor receives in future rounds. Axelrod's (1984) label "shadow of the future" aptly captures this feature.

Assume now that the trustor is conditionally trustful in the following sense: she places trust initially, and when the trustee honors trust in a focal TG, she places trust again in future TGs. Conversely, when the trustee abuses trust, the trustor does not place trust in at least some future games. The trustee can then gain T_2 rather than R_2 in a focal TG by abusing trust. On the other hand, abusing trust will be associated with obtaining only P_2 in (some) future encounters, while honoring trust will result in obtaining $R_2 > P_2$ in those future encounters. Also, the larger the shadow of the future w, the more important are the long-term effects of present behavior. Thus, anticipating that the trustor may place trust conditionally, the trustee has to balance short-term ($T_2 - R_2$) and long-term ($R_2 - P_2$) incentives. For example, if a buyer buys regularly from the same seller, the seller has to consider whether the short-term gain from selling an inferior product to this buyer is worth losing a customer.

This raises the question whether conditional trustfulness can be a basis for rational trust in the sense that the indefinitely repeated TG has an equilibrium such that trust is placed and honored in each round. To answer this question, consider the strategy of the trustor that is associated with the largest rewards for trustworthy behavior of the trustee and with the most severe sanctions for untrustworthy behavior. This is the strategy that prescribes to place trust in the first round and also in future rounds, as long as trust has been placed and honored in all previous rounds. However, as soon as trust is not placed or abused in some round, the trustor refuses to place trust in any future round. Straightforward analysis shows that always honoring trust (and always abusing trust as soon as there has been any deviation from the pattern "place and honor trust") is a strategy maximizing the trustee's payoff for the repeated game against the described strategy of the trustor if and only if

$$w \geq (T_2 - R_2)/(T_2 - P_2)$$

(2)

This condition requires that the shadow of the future is large enough compared to $(T_2–R_2)/(T_2–P_2)$, a convenient measure for the trustee's temptation to abuse trust. Under condition (2), the indefinitely repeated Trust Game has an equilibrium such that trust is always placed and honored.[5] Placing trust conditionally and honoring trust in the sense of the strategies mentioned can be seen as following an informal norm or, respectively, maintaining an informal institution (North 1990) of trustful and trustworthy behavior. In equilibrium, by definition, trustor and trustee have no incentive to deviate from such (p. 315) a norm or institution because each actor maximizes own expected payoffs, given the strategy of the other actor. The informal norm or institution, therefore, is self-enforcing and is not assumed to be exogenously given but results endogenously from equilibrium behavior (see Schotter 1981 and Calvert 1995 for the distinction between institutions as exogenous constraints and as outcomes of equilibrium behavior).

Consider an interpretation of the equilibrium condition that follows the logic for deriving testable hypotheses from game-theoretic models as set forth in section 2.1. Since condition (2) is a necessary and sufficient condition for equilibria in which trust is placed and honored throughout an indefinitely repeated TG, one assumes that placing and honoring trust becomes more likely when condition (2) becomes less restrictive. This leads directly to testable hypotheses on control effects through dyadic embeddedness. Specifically, one would expect that the likelihood of placing and honoring trust increases in the shadow of the future *w* and decreases in the temptation $(T_2–R_2)/(T_2–P_2)$ for the trustee. For our buyer-seller example, this implies that trust is more likely to emerge if the seller is more likely to stay in business and if the temptation for the seller to sell an inferior product is not too large. In terms of Coleman's diagram, condition (2), similar to condition (1), can be seen as a macro-condition characterizing interdependencies between the actors. We have thus again generated predictions on how a macro-condition affects trustfulness and trustworthiness as well as a prediction for a macro-level regularity, represented by Arrow 4 in Coleman's diagram. Namely, the game-theoretic model suggests that the mutually beneficial macro-outcome associated with placed and honored trust becomes more likely when condition (2) becomes less restrictive.

In this game-theoretic model with dyadic embeddedness, trust is purely based on control opportunities of the trustor. It is a game with *complete information*: each actor is informed on the behavioral alternatives and incentives of both actors. Specifically, the trustor knows the behavioral alternatives and the incentives of the trustee. Hence, there is no need—and no opportunity—for the trustor to learn during the game about unobservable characteristics of the trustee. Therefore, this model does not yield hypotheses on learning effects of embeddedness.

Hypotheses on control as well as learning effects can be derived from models of trust games with *incomplete information*. Typically, these are models of finitely repeated games (see Bower, Garber, and Watson 1997; Buskens 2003; Camerer and Weigelt 1988; Dasgupta 1988; Neral and Ochs 1992). Consider the TGI in Figure 14.2. After the initial move by Nature, there are N rather than only one interactions between trustor and trustee. Now, if the trustor places trust in a round that is not the final round of the repeated game, a rational trustee may honor trust for two different reasons. First, the trustee has no incentive to abuse trust at all ($T_2^* < R_2$). Second, the trustee follows an incentive for reputation building although $T_2 > R_2$. The trustee knows that after abusing trust, the trustor can infer that the trustee has payoffs $T_2 > R_2$ and will thus never place trust again. On the other hand, if the trustee honors trust, the trustor remains uncertain about the trustee's incentives and may place trust again in the future. The trustor can anticipate on such behavior of the trustee and may therefore indeed place trust. In this game, the trustor can control the trustee in that placing trust in future (p. 316) rounds depends on honoring trust in the current round—trustfulness is thus again conditional—and the trustor can learn about the incentives of the trustee from the trustee's behavior in previous rounds. The result is a subtle interplay of a trustor who tries to learn about and to control the trustee, taking the trustee's incentives for reputation building into account, and a trustee who balances the long-term effects of his reputation and the short-term incentives for abusing trust, taking into account that the trustor anticipates this balancing.

It can be shown that for large enough N the game has a sequential equilibrium (Kreps and Wilson 1982) that involves placing and honoring trust in some of the N rounds. In this equilibrium, the game starts with trust being placed and honored in some rounds. Afterward, a second phase follows in which the trustor and the trustee with $T_2 > R_2$ randomize their behavior until the trustor does not place trust or the trustee abuses trust. After trust has not been placed or has been abused for the first time, the third and last phase starts in which no trust is placed until the end of the game. A remarkable feature of the model is that, due to reputation effects, much trustfulness and trustworthiness can be induced by equilibrium behavior even if the probability π of a reliable trustee is very small. In the sequential equilibrium, rational learning occurs in the sense that the trustor rationally updates her belief about the probability that she is playing with a reliable trustee. The first phase of the game with trust being placed and honored is shorter, the higher the risk $(P_1 - S_1)/(R_1 - S_1)$ for the trustor, the smaller the number of rounds of the repeated game, and the smaller the probability π. With respect to learning effects of dyadic embeddedness, we thus obtain the hypotheses that the likelihood of placing and honoring trust decreases if the trustor's risk is higher and increases if the trustor's previous experiences with the trustee have been positive (the trustee honored trust) rather than negative (the trustee abused trust).

Network embeddedness

Network embeddedness refers to the case that the interaction of trustor and trustee in the focal trust game is related to their interactions with third parties. As an example, consider reputation systems on online markets. A reputation score provides information for the buyer about previous behavior and performance of the seller. Also, by providing feedback, the buyer can affect the seller's reputation score for future transactions. More generally, network embeddedness, such as ties between trustors that allow for exchange of information about the trustee, provides for additional control of the trustee since his behavior in a focal trust game can have repercussions not only for future behavior of the trustor but also for future behavior of other trustors. Likewise, such network embeddedness implies that the trustor in the focal trust game can use her own experience from previous trust games with the trustee for learning about the trustee, and in addition she can draw on information she receives from other trustors about their experiences.

(p. 317) Network embeddedness can be a substitute for dyadic embeddedness. Assume that the trustee interacts with a different trustor in each round of an indefinitely repeated TG (see Kreps 1990, 106–108). Thus, each trustor plays only once with the trustee. Dyadic embeddedness is then removed completely from the repeated game and replaced by network embeddedness. However, if the trustor in a given round is reliably informed about what has happened in previous rounds, each trustor can condition her behavior in a given round in the same way as a trustor who plays in each round: trust is placed if and only if trust has been honored in all previous rounds. Evidently, the trustee maximizes his payoffs against such behavior of the trustors by honoring trust in each round if condition (2) is fulfilled. Conversely, placing trust is then payoff-maximizing for the trustors under condition (2). Hence, we see that network embeddedness can induce trust that is backed up by a self-enforcing norm.

Alternatively, network embeddedness can complement dyadic embeddedness. The following is an example of a game-theoretic model of trust games that combines dyadic as well as network embeddedness (Buskens and Weesie 2000; see Raub and Weesie 1990 for a related model of network embeddedness for the Prisoner's Dilemma). A trustee interacts with a trustor in an indefinitely repeated TG. After the interaction with a given trustor ends, the trustee goes on playing with another trustor, while information on behavior in the games with the first trustor is communicated to the second trustor with some probability. Interactions with a third trustor start after the interactions with the second trustor have ended, and so forth. One can then study equilibria such that trustors place trust if T_2 is not "too large" and if they do not have information that trust has ever been abused (conditional trustfulness). In addition to our earlier hypotheses on how the likelihood of trust is affected by the shadow of the future and the short-term incentives of the trustee, such models allow for deriving hypotheses on effects of network characteristics. Specifically, the likelihood of placing and honoring trust increases in the density of the network of trustors and in the probability that a trustor transmits information to the next trustor interacting with the trustee. These effects are due to increasing the sanction possibilities of the trustor and, thus, her opportunities for controlling the trustee.

Including learning in models of network effects leads once again to more complex games with incomplete information (see Yamagishi and Yamagishi 1994, 138–139 for a discussion of learning and control effects through network embeddedness that uses other than game-theoretic approaches). Such a model has been provided by Buskens (2003). The trustee plays the TGI in Figure 14.2 with two different trustors. With some probability, each trustor can inform the other trustor on the trustee's previous behavior. If each trustor transmits information to and receives information on the trustee from the other trustor with sufficiently high probability, the first phase of the sequential equilibrium during which trust is placed and honored lasts for more rounds of the repeated game, and in this sense network embeddedness increases trust. Also with information sharing, the length of this phase decreases in the trustor's risk $(P_1 - S_1)/(R_1 - S_1)$, yielding again the hypothesis that trust decreases in the trustor's risk. In the second phase of the equilibrium in which trustor and trustee randomize, the trustor's (p. 318) estimation of the probability that the trustee is of the reliable type increases with positive information about the trustee's past behavior. With network embeddedness, not only own positive experiences with the trustee but also positive information from the other trustor can induce such learning effects. Therefore, we again hypothesize that positive information increases the likelihood of placing trust.

Our models of network embeddedness include the assumption that information trustors receive is reliable and incentive problems associated with the supply of information are neglected (see, e.g., Buskens 2002, 18–20; Raub and Weesie 1990, 648; Williamson 1996, 153–155). Supplying information on the trustee's behavior is a contribution to a public good, namely, enforcing trustworthy behavior of the trustee. With costly contributions, public good production is problematic and contributions cannot be taken for granted (this is a major issue of reputation systems such as eBay's feedback forum; see, e.g., Diekmann et al. 2014). Furthermore, information from third parties can be inconsistent with one's own experiences. Also, information from third parties can be problematic due to misunderstanding or strategic misrepresentation: imagine that trustors are competitors who purchase the same goods from the same seller. In a nutshell, one would expect that effects of network embeddedness are attenuated when such problems become more serious.

Summarizing, we have by now seen how game-theoretic models can be used to generate hypotheses on effects of dyadic and network embeddedness in trust games. A common feature of these models is that they allow to derive when trustfulness and trustworthiness are outcomes of equilibrium behavior, with equilibrium behavior being based on conditional trust. In this sense, the models show that informal norms and institutions of conditional trust can be self-enforcing. Another useful feature of the models is that they not only show that equilibrium behavior in trust games depends on whether these games are isolated encounters or are embedded. Rather, the models also allow for theoretically disentangling different kinds of embeddedness effects: they allow distinguishing effects due to dyadic embeddedness from those due to network embeddedness as well as distinguishing between control and learning effects.

Endogenous embeddedness

A common feature of the game-theoretic models reviewed above is the assumption that embeddedness is exogenously given. We have seen that embeddedness can provide mutual benefits for trustor and trustee. Employing another terminology, embeddedness constitutes "social capital" for trustor and trustee, since embeddedness refers to relations between actors that help to achieve ends—trustfulness and trustworthiness—that could not be achieved without embeddedness (Coleman 1988). For this reason, trustor and trustee may also have incentives to invest in embeddedness.

Until now, there are only a few game-theoretic models that include such investments and thus endogenize embeddedness by simultaneously modeling investments (p. 319) in and effects of embeddedness in trust games. One class of such models assumes indefinitely repeated TGs (Raub, Buskens, and Frey 2013) or finitely repeated TGIs (Frey 2016; Frey, Buskens, and Raub 2015), such that one and the same trustee plays with different trustors. While dyadic embeddedness is exogenously given—the trustee plays repeatedly with each trustor—network embeddedness is endogenous. Before playing repeated trust games, actors—trustors or trustee—can invest in setting up a network between the trustors that allows for information exchange about the behavior of the trustee. These models provide predictions on the effects of network embeddedness on trust like those sketched in section 2.3. The interesting additional feature refers to predictions on investments in network embeddedness. The core result, robust to assuming either indefinitely repeated TGs or finitely repeated TGIs, is an inverse U-shaped relation between the size of a trust problem and incentives for investments. Roughly, incentives to invest in establishing network embeddedness are small for trust problems that are small and can be mitigated through dyadic embeddedness alone. Such incentives are likewise small if trust problems are very large and trustfulness and trustworthiness are unattainable, even if network embeddedness and dyadic embeddedness complement each other. Incentives for investments in network embeddedness are large for trust problems of an intermediate size such that the effects of network embeddedness in addition to dyadic embeddedness make a difference for the behavior of trustors and trustee. For trust problems of intermediate size, investments in network embeddedness provide the conditions for the subsequent emergence of self-enforcing informal norms and institutions of conditional trust since the benefits from conditional trust compared to no trust exceed the costs of the investments. Imagine once again the buyer-seller example. If a buyer buys a small and more or less standard product such as bread, she typically does not ask friends which seller can be expected to sell the best bread. If the buyer is going to buy a house, the deal will only be settled after extensive contracting, and asking friends about the seller will usually not change much. However, when it comes to transactions of an intermediate size and complexity such as buying a secondhand car, information about past performance of the seller seems to have a relevant impact on trust in the seller.

It is important to grasp that these theoretical results apply for trustors as well as trustees. All actors are better off when trust is placed and honored compared to no trust placed. Thus, not only trustors but also trustees themselves may wish to invest in a network that allows for information exchange about the trustee. While network embeddedness makes it less attractive for the trustee to abuse trust in the short run, since the long-run costs of abusing trust increase, it is precisely this feature that may induce trustors to place trust in the first place. Therefore, it can be equilibrium behavior that the trustee himself invests in setting up an information network for trustors and in this way credibly commits himself to honoring trust (see Schelling 1960; Williamson 1985 on the use of commitments in social and economic interactions). For example, not only buyers but also sellers in online markets have an interest in the availability of a reputation system.

(p. 320) One could develop models in which other features of the interaction between trustor and trustee are endogenized. Raub (2016) offers a simple model with endogenous dyadic embeddedness in the sense that actors have a choice whether to invest in the possibility or likelihood that interactions with their partner are repeated. There are some models on commitments (e.g., Raub 2004; Snijders 1996; Weesie and Raub 1996) providing opportunities for trustees to modify their own future incentives or, as Coleman (1990) put it, to construct their social environment. Because these models are mostly applied to isolated encounters, we do not discuss them further.

Experimental Evidence from Embedded Trust Games

We first discuss experimental evidence on the effects of dyadic embeddedness and continue with evidence on network embeddedness and endogenous embeddedness.[6] Few experiments on embeddedness effects employ the IG. We mainly consider experiments with TGs and TGIs. The early experiments by Camerer and Weigelt (1988) show that subjects may behave similarly in the finitely repeated TG and finitely repeated TGI. In their experiment, behavioral patterns were similar in conditions in which it was common knowledge like in the TGI that a proportion of subjects in the role of trustees have no material incentive to abuse trust as in conditions like in the TG without such trustees. Camerer and Weigelt's explanation was that even if all trustees have material incentives to abuse trust, subjects anticipate that there is a proportion of subjects who have other-regarding preferences that prevent them from abusing trust even in the last round of a finitely repeated trust game. This anticipated proportion of intrisically reliable trustees implies that any experimentally played series of standard TGs can be interpreted theoretically as a repeated TGI. Therefore, we will not always distinguish explicitly between TGs and TGIs.

Effects of dyadic embeddedness

Camerer and Weigelt (1988) were the first to test explicitly behavioral patterns predicted by the sequential equilibrium in finitely repeated TGIs. Neral and Ochs (1992), Anderhub, Engelmann, and Güth (2002), and Brandts and Figueras (2003) did various follow-up experiments. Camerer (2003, 446–453) provides a detailed overview of these experiments. Experiments confirm the general pattern predicted, namely, that trustfulness and trustworthiness are high in early rounds and decrease when the end of the repeated game approaches (dyadic control). Trustfulness is largely conditional on past trustworthiness and almost absent after any abuse of trust (dyadic learning). The trustworthiness of so-called unreliable trustees is largely a strategic response to (p. 321) the anticipated conditional behavior of trustors and reaches very low levels in the last round of a repeated game (e.g., Anderhub, Engelmann, and Güth 2002). Furthermore, as predicted by the theory, Brandts and Figueras (2003) find that trustfulness and trustworthiness increase with the probability that a trustee is reliable; and Anderhub, Engelmann, and Güth (2002) find that behavior in the last rounds does indeed not vary significantly with the total length of the repeated game. Summarizing, the sequential equilibrium described for the finitely repeated TGIs predicts quite some behavioral patterns reasonably well. Still, the experiments of Neral and Ochs (1992), Anderhub, Engelmann, and Güth (2002), and Brandts and Figueras (2003) also show that behavior of subjects does not follow the predicted patterns in all respects. For example, the theoretical model implies that in the second phase of the game in which trustors and unreliable trustees randomize, the probability that trustors place trust increases (!) with the trustee’s temptation to abuse trust, while the trustee’s behavior is not expected to depend on his payoffs. This implication is not only counterintuitive but also inconsistent with experimental findings (see Neral and Ochs 1992). The results of Anderhub, Engelmann, and Güth furthermore indicate that subjects do not randomize in the second phase as predicted, but use heuristics that appear like stopping rules, prescribing that they deterministically place or honor trust until a specific number of rounds is left and then stop trusting.

Results from some *other experiments* are quite in line with these findings. Gautschi (2000) reports conditional trustfulness for two and three times repeated TGs and dyadic control effects in the sense that trust increases with the number of remaining rounds. Kollock (1994) finds similar effects in a contextualized trust setting with buyers and sellers. A difference with the other studies is that Gautschi and Kollock find more untrustworthy behavior in early rounds of the games. Most likely, this is related to Gautschi and Kollock letting subjects play relatively few repeated games, while Camerer and Weigelt let subjects play the finitely repeated TGI many times. That behavioral patterns in finitely repeated trust games approach the patterns predicted by the sequential equilibrium, especially after subjects gained experience, is documented by Camerer and Weigelt (1988), Brandts and Figueras (2003), and Van Miltenburg, Buskens, and Raub (2012). Bohnet, Harmgart, Huck, and Tyran (2005) furthermore show that rates of honored trust are higher in early rounds if trustees can observe the behavior of other trustees, which indicates that some trustees have to learn to invest in a good reputation from observing other trustees' behavior.

Engle-Warnick and Slonim (2004, 2006) compare *finitely* and *indefinitely repeated games*. In principle, the trustor's opportunities to exercise control in an indefinitely repeated game with constant continuation probability are the same in round t and in round $t + 1$. Still, the authors find decreasing trust over time in such games. However, this decrease is much smoother than in the finitely repeated games. This can be interpreted as learning effects related to negative experiences reducing trust over time, and subsequently trust is difficult to restore. On the other hand, trust remains reasonably high because control opportunities do not diminish over time and enable some pairs to continue to trust each other. An additional explanation for decreasing trust in (p. 322) indefinitely repeated games might be that subjects believe that after many rounds the probability increases that a specific round is the last one, even if experimenters do their very best to show that the continuation probability is constant (e.g., by using a publicly thrown die).

While there are many experiments on the Investment Game (IG), only a few use repeated IGs. Cochard, Nguyen-Van, and Willinger (2004) find results on finitely repeated IGs in line with empirical regularities found for the TG and TGI. Trustors send more in the IG if there is a longer future (dyadic control), but if trustees do not return enough they stop sending (dyadic learning). In early rounds, trustors send more if trustees returned more. While Cochard, Nguyen-Van, and Willinger refer to this finding as a reciprocity effect, it can also be interpreted as a learning effect. Again, there is a strong endgame effect, although it is observed very late in the games.

Dubois, Willinger, and Blayac (2012) found mixed support for the effects of dyadic embeddedness in a finitely repeated IG. In their experiment, subjects played in groups of six and in every period were paired randomly to play an IG. Compared to a treatment in which players cannot identify one another over periods, stable trustee identities lead to more trustworthiness but not to more trustfulness. Trustfulness only increased when trustors were also identifiable over the periods. While the game-theoretic arguments sketched in section 2.2 do not suggest that trustor identifiability matters, Dubois, Willinger, and Blayac find that two-sided identifiability allows for the emergence of bilateral trust-reciprocity in which trustfulness increases over time. Altogether, the evidence on dyadic embeddedness shows that informal norms of placing and honoring trust can emerge if the "shadow of the future" is large enough, although the success might depend on how the repeated interaction is exactly institutionalized.

Effects of network embeddedness

Experiments with trust games that include network embeddedness are still scarce. Bolton, Katok, and Ockenfels (2004; see also Bolton and Ockenfels 2009) compare one-shot TGs that are isolated encounters in the strict sense, finitely repeated TGs with the same partner, and a third treatment in which subjects play multiple one-shot TGs with different partners but obtain information about the past behavior of their partners in interactions with other subjects (for a similar setup and results, see Bohnet and Huck 2004). In the one-shot TGs, trustfulness and trustworthiness decline quickly after subjects have some experience. Trustfulness and trustworthiness remain high in the repeated TGs and collapse only in the last couple of rounds. This finding resonates with evidence on effects of dyadic embeddedness. In the third treatment with network embeddedness, there is initially less trustfulness and trustworthiness than in the finitely repeated TG setting, but trustees apparently learn fast enough that they have a problem if they do not honor trust. In this treatment, trustfulness and trustworthiness stabilize for some time in the middle of the series of interactions, although at a (p. 323) somewhat lower level than if the TG is repeated between the same two partners. This suggests that network embeddedness is an imperfect substitute for repeated personal interaction (dyadic embeddedness). Finally, as in the repeated TG setting, trust collapses in the last rounds also in this treatment.

Bolton, Katok, and Ockenfels (2004) interpret their third treatment as an experimental implementation of a reputation system that is common for online transactions. The treatment could also be interpreted as a complete network in which information diffusion is perfect. Huck, Lünser, and Tyran (2010) varied network embeddedness more gradually. They had groups of four trustors and four trustees interact over thirty rounds. Every round, trustors and trustees were paired randomly to play a TG. In one treatment, trustors knew only their private history with trustees. In a second treatment, the trustors were located on a circle and knew also the history of their right-hand neighbor. In a third treatment, trustors knew the entire history. In accordance with theoretical predictions, trustfulness and trustworthiness increased with the level of information sharing, although the increase was not significant between all adjacent treatments. Somewhat surprisingly, Huck, Lünser, and Tyran (2012) find hardly any effect of network embeddedness in a highly similar experiment. Note that the Huck, Lünser, and Tyran (2010) and the Bolton, Katok, and Ockenfels (2004) reputation treatments involve opportunities for learning as well as control through third parties. While indicating that network embeddedness matters, these experiments leave open through which mechanism—learning or control or both—network embeddedness promotes trust.

Buskens, Raub, and Van der Veer (2010) introduce a network setting with subjects playing finitely repeated TGs in groups of three. There are two trustors who take turns in playing with the same trustee. The design varies network embeddedness, namely, whether or not a trustor obtains information about the interactions the other trustor has with the trustee (see Barrera and Buskens 2009 for a related study on the IG). As expected, there is more trust in the condition with network embeddedness. Similar to findings from other experiments, within dyads, trustors are more trustful after positive experiences with the trustee, and trustfulness as well as trustworthiness collapse near the end of the game. This is once again evidence for dyadic control as well as dyadic learning. The theoretical analysis of Buskens (2003) implies that the decrease in trustfulness and trustworthiness should start later with network embeddedness because of the network control effect, that is, because all actors know that a *single* abuse of trust can lead *both* trustors to stop placing trust. Buskens, Raub, and Van der Veer (2010) do not find evidence for this network control effect on the trustfulness of the trustors: the increase in trustfulness is purely based on network learning. Still, they do find evidence for network control effects on trustworthiness of the trustee. They offer a bounded rationality argument for why network control has an effect only for trustees and not for trustors: the trustee needs to anticipate third-party sanctions, while the trustor needs a further step of strategic reasoning—namely, anticipating that the trustee anticipates the third-party sanctions. One might expect that this bounded rationality argument disappears when subjects gain more experience with the game, **(p. 324)** but this is not confirmed in a setup in which subjects play more replications of the repeated game (Van Miltenburg, Buskens, and Raub 2012). Cassar and Rigdon (2011) investigate trust in three actor-networks in the IG, but they apply one-shot interactions. Therefore, the type of learning effects they consider differs from the type of embeddedness effects we considered here.

Summarizing, the experiments on network embeddedness show that networks can be considered informal institutions that enable norms for placing and honoring trust and that complement or substitute effects of dyadic embeddedness. Still, the precise mechanisms that produce the effects of network embeddedness are not completely in line with the mechanisms predicted by the game-theoretic models.

Emergence and effects of endogenous network embeddedness

Experiments with trust games that include endogenous network embeddedness are even more scarce than trust experiments with exogenous network embeddedness. Frey, Buskens, and Corten (2016) study investments in and effects of endogenous network embeddedness in an experimental setup in which two trustors interact a finite number of times with the same trustee in TGIs. Before the TGIs are played, the trustee is privately told his type and there is an opportunity to invest some "points" to establish network embeddedness—information exchange between the trustors about the behavior of the trustee. The experiment replicates the finding that network embeddedness facilitates trust (see section 3.2). Moverover, the substantial levels of investments in establishing network embeddedness by trustors as well as trustees confirm that actors may invest in network embeddedness in the expectation that this benefits them in the trust interactions. Frey, Buskens, and Corten also tested the hypothesis of an inverse U-shaped relation between the size of the trust problem and investments in and effects of network embeddedness (see section 2.4) but found no support for this hypothesis. Specifically, network effects were not diminished, and investments in network embeddedness were not less frequent, in experimental conditions in which the trust problem was very large (low probability π of a reliable trustee) or very small (large π). This lack of support should be interpreted with some caution because subject behavior turned out to be generally rather insensitive to the manipulation of π. A final noteworthy finding of Frey, Buskens, and Corten is that network embeddedness tends to promote trust more strongly if it is established endogenously rather than imposed exogenously. This could be due to self-selection: actors who are sensitive to network embeddedness are more likely to establish it. Also, the result could be due to costly signaling: a trustee's costly investment in establishing embeddedness credibly signals that he has no intention to abuse trust (see Frey, forthcoming, for a theoretical model of investments in network embeddedness as costly signals of trustworthiness).

(p. 325) In the Frey, Buskens, and Corten experiment, participants took explicit networking decisions, but once a link for information sharing between the trustors was established, information was exchanged automatically. Another approach to the study of the emergence and effects of information networks is to focus on the actual use of potentially costly opportunities to pass on or request information. Experiments following this approach have also shed some light on the conditions under which actors will endogenously exchange information. In the study of Abraham, Grimm, Ness, and Seebauer (2014), subjects interacted in groups of four trustors and four trustees over a finite number of periods. Every period, trustors and trustees were first randomly matched to play an IG, and then the trustors could pass on information about their transactions. In a control treatment in which trustors could not pass on information, trustfulness and trustworthiness were significantly lower than in various treatments with a possibility for information sharing, including treatments with costly information sharing. Trustors shared information most frequently when shared information was made available to all trustors rather than to only one randomly chosen trustor and when sharing information was free of costs. Gerxhäni, Brandts, and Schram (2013) study the use and effects of information sharing in experimental labor markets, representing employer-employee interactions in a game that differs somewhat from the trust games described in section 2. Their findings indicate that an employer may be willing to answer a request for information about a job candidate's past performance by another employer even if this is costly, motivated by the expectation that the benefiting employer will later reciprocate the favor. In treatments in which anonymity prevented direct reciprocity, information sharing was also frequent and probably motivated by employers wanting to contribute their share to the provision of a collective good—namely, incentives for trustworthiness—and expecting that others will only contribute their share so long as contribution levels are sufficiently high. Finally, in an experiment designed to represent the situation of money lenders, Brown and Zehnder (2007) show that trustors (lenders) may contribute to institutions for information sharing about clients even if they are in competition with some of those who could benefit from the information they contribute.

It has been mentioned that other elements of the interaction between trustors and trustees can be endogenized, too, in theoretical models. Empirically, Kollock (1994) studied the role of trust in the formation of long-term relations in a contextualized trust experiment with buyers and sellers and showed that buyers were more likely to form long-term relations (established dyadic embeddedness endogenously) if sellers could misrepresent product quality. Frey and Van de Rijt (2016) show in an abstract experiment with trust games that subjects are more likely to choose the same trustee after honored trust than in a setting that lacks the trust problem. Findings on the conditions under which commitments or other signals can induce trust can be found, among others, in Snijders (1996), Snijders and Buskens (2001), Bolle and Kaehler (2007) for one-shot games, and for repeated games in Przepiorka and Diekmann (2013).

(p. 326) Conclusions

We have shown alternative versions of trust games and also described how, employing game-theoretic tools and experimental research, they can be used to shed light on trust as explanandum. More specifically, we have highlighted that studying trust games can help to understand how macro-conditions—various forms of embeddedness—affect trust by providing a basis for self-enforcing informal norms and institutions of conditional trust. We have likewise shown that game-theoretic equilibrium behavior can induce actors to actively modify the conditions under which they subsequently interact in trust games so that informal norms and institutions of conditional trust become self-enforcing. While there is quite some research on effects of exogeneous embeddedness, endogenous embeddedness is less well explored theoretically and experiments with endogenous embeddedness are still scarce.

Experimental evidence shows that, not surprisingly in light of what is common in social science, quantitative point predictions from game-theoretic models fail. Alternatively, we sketched, using Coleman's diagram, how game-theoretic models can be employed for generating qualitative hypotheses on embeddedness effects, based on comparative statics. Such hypotheses on changes "at the margin" often succeed in predicting the signs of coefficients. Still, there is room for improving the accuracy of such predictions. In general, assuming game-theoretic rationality together with the selfishness assumption of "utility = own money" seems problematic, not only in the sense of being incompatible with experimental evidence for quite some trust in isolated encounters but also in leading one to expect considerably more trustfulness and trustworthiness in embedded settings than experimental evidence shows (see Bolton and Ockenfels 2009 for a similar observation based on evidence from research on reputation systems for online markets).

Note, too, that our review focused on experimental evidence on trust games. Experimental designs have advantages, but for establishing the robustness of empirical findings and regularities it seems quite useful to employ complementary empirical designs and data, that is, to use experiments as well as, say, survey designs, archival data, and vignette studies for testing the same sets of hypotheses on trust as an explanandum (see Buskens and Raub 2013 for further discusison).

From the theoretical end, we have throughout focused on game-theoretic models. Many assumptions in standard game-theoretic models can be seen as problematic. This is the case for the selfishness assumption and likewise for various rationality assumptions in such models. When one wishes to come up with superior models, the challenge is to satisfy a number of criteria simultaneously. Namely, one has to replace problematic assumptions of standard game theory so that the alternatives provide better accounts for the overall patterns of empirical evidence on trust games rather than exclusively accounting for some specific empirical "anomaly" relative to (p. 327) game-theoretic predictions. Furthermore, one would need an alternative that, like game theory, models a core feature of trust games, namely, interdependence between trustor and trustee, and comes up with assumptions on how behavior in trust games is affected by such interdependency. Alternatives such as models including assumptions on other-regarding preferences or bounded rationality assumptions, including pure learning models, and other models from behavioral game theory go some way in this direction, but it is hard to overlook that such models are typically tailor-made for specific applications only. Given this, it seems certainly worthwhile to bet on theoretical pluralism and foster the development of theoretical alternatives rather than to prematurely dismiss alternatives.

Acknowledgments

The order of authorship is alphabetical. Support was provided by NWO (PIONIER-program “The Management of Matches,” grants S 96–168 and PGS 50–370 for Raub, NWO Graduate Program 2008/2009 for ICS and Buskens, Raub, and Frey) and by Utrecht University (funding for the Strategic Theme “Institutions for Open Societies”). Raub acknowledges the hospitality of Nuffield College, University of Oxford. We thank Nynke van Miltenburg for assistance with collecting literature used in this chapter.

References

Abraham, M., V. Grimm, C. Ness, and M. Seebauer. 2014. Reputation formation in economic transactions. *Journal of Economic Behavior & Organization* 121: 1–14.

Anderhub, V., D. Engelmann, and W. Güth. 2002. An experimental study of the repeated trust game with incomplete information. *Journal of Economic Behavior & Organization* 48: 197–216.

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

Axelrod, R. 1984. *The Evolution of Cooperation*. New York: Basic Books.

Barber, B. 1983. *The Logic and Limits of Trust*. New Brunswick, NJ: Rutgers University Press.

Barrera, D., and V. Buskens. 2009. Third-party effects. In K. S. Cook, C. Snijders, V. Buskens, and C. Cheshire, eds., *eTrust: Forming Relationships in the Online World*, 37–72. New York: Russell Sage.

Berg, J., J. Dickhaut, and K. McCabe. 1995. Trust, reciprocity, and social history. *Games and Economic Behavior* 10: 122–142.

Binmore, K. 1998. *Game Theory and the Social Contract. Vol. 2, Just Playing*. Cambridge, MA: MIT Press.

Bohnet, I., H. Harmgart, S. Huck, and J.-R. Tyran. 2005. Learning trust. *Journal of the European Economic Association* 3: 322–329.

Bohnet, I., and S. Huck. 2004. Repetition and reputation: Implications for trust and trustworthiness when institutions change. *American Economic Review Papers and Proceedings* 94: 362–366.

Bolle, F., and J. Kaehler. 2007. Introducing a signaling institution: An experimental investigation. *Journal of Institutional and Theoretical Economics* 163: 428–447.

Bolton, G. E., E. Katok, and A. Ockenfels. 2004. How effective are online reputation mechanisms? An experimental study. *Management Science* 50: 1587–1602.

Bolton, G. E., and A. Ockenfels. 2009. The limits of trust in economic transactions: Investigations of perfect reputation systems. In K. S. Cook, C. Snijders, V. Buskens, and C. Cheshire, eds., *eTrust: Forming Relationships in the Online World*, 15–36. New York: Russell Sage.

Bower, A., S. Garber, and J. C. Watson. 1997. Learning about a population of agents and the evolution of trust and cooperation. *International Journal of Industrial Organization* 15: 165–190.

Brandts, J., and N. Figueras. 2003. An exploration of reputation formation in experimental games. *Journal of Economic Behavior & Organization* 50: 89–115.

Brown, M., and C. Zehnder. 2007. Credit reporting, relationship banking, and loan repayment. *Journal of Money, Credit and Banking* 39: 1883–1918.

Buskens, V. 2002. *Social Networks and Trust*. Boston, MA: Kluwer.

Buskens, V. 2003. Trust in triads: Effects of exit, control, and learning. *Games and Economic Behavior* 42: 235–252.

Buskens, V., and W. Raub. 2002. Embedded trust: Control and learning. *Advances in Group Processes* 19: 167–202.

Buskens, V., and W. Raub. 2013. Rational choice research on social dilemmas. In R. Wittek, T. A. B. Snijders, and V. Nee, eds., *Handbook of Rational Choice Social Research*, 113–150. Stanford, CA: Stanford University Press.

Buskens, V., W. Raub, and J. van der Veer. 2010. Trust in triads: An experimental study. *Social Networks* 32: 301–312.

Buskens, V., and J. Weesie. 2000. Cooperation via networks. *Analyse & Kritik* 22: 44–74.

(p. 329) Calvert, R. L. 1995. Rational actors, equilibrium, and social institutions. In J. Knight and I. Sened, eds., *Explaining Social Institutions*, 57–95. Ann Arbor: University of Michigan Press.

Camerer, C. F. 2003. *Behavioral Game Theory: Experiments in Strategic Interaction*. New York: Russell Sage.

Camerer, C. F., and K. Weigelt. 1988. Experimental tests of a sequential equilibrium reputation model. *Econometrica* 56: 1–36.

Cassar, A., and M. Rigdon. 2011. Trust and trustworthiness in networked exchange. *Games and Economic Behavior* 71: 282–303.

Charness, G., and V. Shmidov. 2013. Trust and reciprocity. *Foundations and Trends in Microeconomics* 10: 167–207.

Cochard, F., P. Nguyen-Van, and M. Willinger. 2004. Trusting behavior in a repeated investment game. *Journal of Economic Behavior & Organization* 55: 31–44.

Coleman, J. S. 1988. Social capital in the creation of human capital. *American Journal of Sociology* 94: S95–S120.

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

Cooper, D. J., and J. H. Kagel. 2015. Other-regarding preferences: A selective survey of experimental results. In J. H. Kagel and A. E. Roth, eds., *Handbook of Experimental Economics*, vol. 2, 217–289. Princeton, NJ: Princeton University Press.

Craswell, R. 1993. On the uses of “trust.” *Journal of Law and Economics* 36: 487–500.

Dasgupta, P. 1988. Trust as a commodity. In D. Gambetta. ed., *Trust: Making and Breaking Cooperative Relations*, 49–72. Oxford: Blackwell.

Diekmann, A., B. Jann, W. Przepiorka, and S. Wehrli. 2014. Reputation formation and the evolution of cooperation in anonymous online markets. *American Sociological Review* 79: 65–85.

Dubois, D., M. Willinger, and T. Blayac. 2012. Does players’ identification affect trust and reciprocity in the lab?*Journal of Economic Psychology* 33: 303–317.

Engle-Warnick, J., and R. L. Slonim. 2004. The evolution of strategies in a repeated trust game. *Journal of Economic Behavior & Organization* 55: 553–573.

Engle-Warnick, J., and R. L. Slonim. 2006. Learning to trust in indefinitely repeated games. *Games and Economic Behavior* 54: 95–114.

Fehr, E., and H. Gintis. 2007. Human motivation and social cooperation: Experimental and analytical foundations. *Annual Review of Sociology* 33: 43–64.

Fehr, E., and K. M. Schmidt. 1999. A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics* 114: 817–868.

Frey, V., V. Buskens, and W. Raub. 2015. Embedding trust: A game-theoretic model for investments in and returns on network embeddedness. *Journal of Mathematical Sociology* 39: 39–72.

Frey, V. 2016. *Network Formation and Trust*. PhD thesis: Utrecht University.

Frey, V. (forthcoming). Boosting trust by facilitating communication: A model of trustee investments in information sharing. *Rationality and Society*.

Frey, V., V. Buskens, and R. Corten, 2016. Investments in and returns on embeddedness: An experiment with Trust Games. Working Paper, Utrecht University.

Frey, V., and A. van de Rijt. 2016. Arbitrary inequality in reputation systems. *Scientific Reports 6*, 38304.

Fudenberg, D., and E. Maskin. 1986. The folk theorem in repeated games with discounting or with incomplete information. *Econometrica* 54: 533–554.

(p. 330) Gautschi, T. 2000. History effects in social dilemma situations. *Rationality and Society* 12: 131–162.

Gërxhani, K., J. Brandts, and A. Schram. 2013. The emergence of employer information networks in an experimental labor market. *Social Networks* 35: 541–560.

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

Huck, S., G. K. Lünser, and J. R. Tyran. 2010. Consumer networks and firm reputation: A first experimental investigation. *Economics Letters* 108: 242–244.

Huck, S., G. K. Lünser, and J. R. Tyran. 2012. Competition fosters trust. *Games and Economic Behavior* 76: 195–209.

Johnson, N. D., and A. A. Mislin. 2011. Trust games: A meta-analysis. *Journal of Economic Psychology* 32: 865–889.

Kollock, P. 1994. The emergence of exchange structures: An experimental study of uncertainty, commitment, and trust. *American Journal of Sociology* 100: 313–345.

Kreps, D. M. 1990. Corporate culture and economic theory. In J. E. Alt and K. A. Shepsle, eds., *Perspectives on Positive Political Economy*, 90–143. Cambridge: Cambridge University Press.

Kreps, D. M., and R. Wilson. 1982. Sequential equilibria. *Econometrica* 50: 863–894.

Miltenburg, N. van, V. Buskens, and W. Raub. 2012. Trust in triads: Experience effects. *Social Networks* 34: 425–428.

Neral, J., and J. Ochs. 1992. The sequential equilibrium theory of reputation building: A further test. *Econometrica* 60: 1151–1169.

North, D. C. 1990. *Institutions, Institutional Change and Economic Performance*. Cambridge: Cambridge University Press.

Ortmann, A., J. Fitzgerald, and C. Boeing. 2000. Trust, reciprocity, and social history: A re-examination. *Experimental Economics* 3: 81–100.

Przepiorka, W., and A. Diekmann. 2013. Temporal embeddedness and signals of trustworthiness: Experimental tests of a game theoretic model in the United Kingdom, Russia, and Switzerland. *European Sociological Review* 29: 1010–1023.

Rapoport, A. 1974. Prisoner's dilemma—Recollections and observations. In A. Rapoport, ed., *Game Theory as a Theory of Conflict Resolution*, 17–34. Dordrecht: Reidel.

Rasmusen, E. 2007. *Games and Information: An Introduction to Game Theory*. 4th ed. Oxford: Blackwell.

Raub, W. 2004. Hostage posting as a mechanism of trust: Binding, compensation, and signaling. *Rationality and Society* 16: 319–366.

Raub, W. 2016. *Strategic Tie Formation*. Mimeo.

Raub, W., V. Buskens, and M. A. L. M. van Assen. 2011. Micro-macro links and microfoundations in sociology. *Journal of Mathematical Sociology* 35: 1–25.

Raub, W., V. Buskens, and V. Frey. 2013. The rationality of social structure: Cooperation in social dilemmas through investments in and returns on social capital. *Social Networks* 35: 720–732.

Raub, W., and J. Weesie. 1990. Reputation and efficiency in social interactions: An example of network effects. *American Journal of Sociology* 96: 626–654.

Schelling, T. C. 1960. *The Strategy of Conflict*. London: Oxford University Press.

Schotter, A. 1981. *The Economic Theory of Social Institutions*. Cambridge: Cambridge University Press.

Snijders, C. 1996. *Trust and Commitments*. Amsterdam: Thesis Publishers.

(p. 331) Snijders, C., and V. Buskens. 2001. How to convince someone that you can be trusted? The role of "hostages." *Journal of Mathematical Sociology* 25: 355–383.

Weesie, J., and W. Raub. 1996. Private ordering: A comparative institutional analysis of hostage games. *Journal of Mathematical Sociology* 21: 201–240.

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

Williamson, O. E. 1996. *The Mechanisms of Governance*. New York: Oxford University Press.

Yamagishi, T., and M. Yamagishi. 1994. Trust and commitment in the United States and Japan. *Motivation and Emotion* 18: 129–166. (p. 332)

Notes:

[1] For readability, "she" refers to the trustor and "he" to the trustee.

[2] Rasmusen (2007) is a textbook on game theory that is accessible to readers with modest training in formal theoretical models. Where necessary, we provide intuition on game-theoretic concepts and assumptions. We refer the reader to Rasmusen's book for further information. Unless explicitly indicated otherwise, we use the standard assumptions sketched here for all games discussed in this article.

[3] If the amount M_1 sent by the trustor is "small," a "small" amount K_2 returned by the trustee could also be interpreted as a punishment the trustee inflicts on the trustor for not trusting the trustee.

[4] Note that we use "trust game" generically for games introduced in this section, while "Trust Game" and TG refer specifically to the game in Figure 14.1.

[5] A very similar result can be obtained for an indefinitely repeated Investment Game and, indeed, for a large class of other repeated games. The equilibrium, however, is not unique. For example, never placing trust, while placed trust would always be abused, is always an equilibrium of the indefinitely repeated game. The "folk theorem" (e.g., Fudenberg and Maskin 1986; Rasmusen 2007, chap. 5.2) for repeated games implies that the indefinitely repeated TG has many other equilibria, too, for large enough w. See Buskens and Raub (2013, 125) for a brief discussion of this issue.

[6] See Buskens and Raub (2013) for additional information on some of the experiments reviewed here as well as on complementary evidence employing survey studies and vignette designs.

Vincent Buskens

Vincent Buskens is Professor of Theoretical Sociology in the Department of Sociology at Utrecht University and the Interuniversity Center for Social Science Theory and Methodology (ICS).

Vincenz Frey

Vincenz Frey is a postdoctoral researcher in the Department of Sociology at Utrecht University and the Interuniversity Center for Social Science Theory and Methodology (ICS).

Werner Raub

Werner Raub is Professor of Sociology at Utrecht University and the Interuniversity Center for Social Science Theory and Methodology (ICS).